RISK OF BIAS ASSESSMENT FOR STUDIES OF EXPOSURES

RISK OF BIAS ASSESSMENT FOR NON-RANDOMIZED STUDIES OF ENVIRONMENTAL EXPOSURES

By Rebecca Lynn Morgan, M.P.H.

A Thesis Submitted to the School of Graduate Studies in Partial Fulfilment of the Requirements for the Degree of Doctor of Philosophy

McMaster University DOCTOR OF PHILOSOPHY (2018)

Hamilton, Ontario (Health Research Methodology Program).

TITLE: Risk of bias assessment for non-randomized studies of environmental exposures.

AUTHOR: Rebecca L. Morgan, M.P.H. (McMaster University)

SUPERVISOR: Professor Holger J. Schünemann, M.D., Ph.D., M.Sc.

NUMBER OF PAGES: xxii, 230

LAY ABSTRACT

When making a decision about interventions to reduce or remove an environmental exposure, evidence is needed to weigh the desirable and undesirable consequences of the decision. No research study is perfect. Most of the studies documenting environmental exposures cannot control for the fact that people who might be highly exposed may have different characteristics compared to those who have low levels of exposure other than just the exposure itself. For example, people exposed to more environmental air pollution living in inner cities may also be more likely to smoke or have occupational exposures that could predispose them to lung cancer than those exposed to lower levels of air pollution. Understanding limitations in studies that address those questions informs our certainty that the data represents the truth. The greater the confidence we have in the data, the more likely we are to be certain that removing or reducing exposure will lead to a desirable outcome. A tool can be used to walk people through the evaluation of limitations within each study. However, it is important that the tool evaluates the correct limitations within the study. It is also important that people using the tool can apply it reliably. Without a reliable or valid tool to evaluate the limitations of the studies, it can be difficult to inform decisions on whether or not to implement specific policies.

In our study, we tested the ability of a new and well-developed tool (ROBINS for interventions) to identify the limitations in studies linking environmental exposures to health outcomes. Based on the findings from our evaluation, we modified our protocol to see if we could improve our ability to evaluate these studies of environmental

exposures. We asked people with an understanding of scientific methods to independently evaluate 35 studies with our modified tool (ROBINS for exposures). We compared those responses to see whether all the reviewers came up with similar decisions and if their decision was similar or different than the conclusion they made using more commonly used tools.

Based on our results, we determined that our modified tool does provide a consistent evaluation of study limitations and accurately measures the limitations present in studies of exposure. This tool can be used to inform decisions about removing or reducing one's exposure to environmental hazards.

ABSTRACT

When using evidence from non-randomized studies (NRS) to answer questions about the effects of environmental exposures on health, it is important to assess risk of bias (RoB) of individual studies as part of determining the certainty in the body of evidence. The recently released RoB in Non-randomized Studies of Interventions (ROBINS-I) instrument has undergone careful development and piloting on NRS of health interventions. A key feature of ROBINS-I is evaluating the RoB of studies against an ideal target trial, therefore establishing a structured comparison of RoB against a reference standard. While several instruments exist to evaluate the RoB of NRS of exposure, none of them use such a structured comparison of RoB. Using the fundamental design of ROBINS-I, we explored development of a version of the instrument to evaluate RoB in studies of environmental exposure. We identified important modifications necessitating a distinct instrument: The RoB instrument for NRS of exposures.

This work highlights the importance of standardized methods for environmental health decision making, proposes a modified instrument to evaluate the RoB of NRS of exposures, provides guidance for the implementation of the instrument and integration into structured evidence-synthesis frameworks (such as GRADE [Grading of Recommendations Assessment, Development and Evaluation]), and presents evidence on the reliability and validity of the instrument. The RoB instrument for NRS of exposures delivers a standardized instrument that systematic review authors and guideline developers can use to evaluate RoB in NRS of exposures. The nature of these

methodological changes allow better integration of RoB assessment in the environmental health field with GRADE.

ACKNOWLEDGMENTS

For my mother (1955 – 2016), who would have reviewed this and other pages for spelling and grammar.

There is no feeling greater than the gratitude that I have for the individuals who supported me during this process. “This process” broadly refers to all years of my academic and non-academic pursuits culminating in this work.

To my supervisor, Holger Schünemann, and committee members, Kris Thayer, Nancy Santesso, and Alison Holloway, I thank you for your guidance, honesty, passion, patience, and understanding. I returned to school for my Ph.D. because I wanted to learn, and you taught. Your mentorship gave me the opportunity to enter a completely new field, explore new topics, tackle new challenges, and meet new colleagues. For me, this has been invaluable.

To my peers and colleagues in the Health Research Methodology program, in the Department of Health Research Methods, Evidence, and Impact, and at McMaster, I thank you for your friendship, support, sincerity, laughter, and comradery. To my mentors, far and wide, I have been extremely fortunate to receive your guidance. Your encouragement and continued support are always needed and very welcome.

To my parents and brother, I have no words powerful enough to describe how integral you are to my entire being. You are my world. I am so very lucky.

To my extended family, you kept me afloat with jokes and tea and hugs. To my tri-pie in crime, thanks for sharing your wisdom and strength.

To my partner and to my friends, in every corner of the globe, you provided the ideal comprehensive care model throughout this process. The well-balanced combination of cat guardians, cooking, adventures, companionship, and advice has sustained me, as has your love, stability, pictures, snail mail, tolerance of presentation rehearsals, and visits.

Rebecca L. Morgan

Hamilton, Ontario – January 2018

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter Four

Chapter Five

LIST OF APPENDICES

Chapter Three

Chapter Four

Chapter Five

LIST OF ABBREVIATIONS

ACROBAT-NRSI: A Cochrane Risk of Bias Tool – for Non-randomized Studies of Interventions

AEC: absolute error coefficient

AHRQ: Agency for Healthcare Research and Quality

ANOVA: analysis of variance

ASTDR: Agency for Toxic Substances and Disease Registry

BMI: Body mass index

BPA: bisphenol-A

CDC: Centers for Disease Control and Prevention

CI: confidence interval

CiE: Certainty in the evidence

CoE: Certainty of evidence

dB: decibel

EFSA: European Food Safety Authority

EPA: Environmental Protection Agency

EtD: Evidence-to-decision

G theory: generalizability theory

GRADE: Grading of Recommendations Assessment, Development, and Evaluation

ICC: intraclass correlation

IRIS: Integrated Risk Information System

NHANES: National Health and Nutrition Examination Survey

NOS: Newcastle-Ottawa Scale

NRC: National Research Council

NRS: non-randomized studies

NTP: National Toxicology Program

OHAT: Office of Health Assessment and Translation

ORoC: Office of the Report on Carcinogen

PBDE: polybrominated diphenyl ethers

PECO: population, exposure, comparator, outcome

PFOA: perfluorooctanoic acid

PICO: population, intervention, comparator, outcome

PM_{10}: particulate matter with aerodynamic diameter less than 10 µm

$PM_{2.5}$: particulate matter with aerodynamic diameter less than 2.5 µm

PME: particulate matter exposure

RCT: randomized controlled trial

RoB: risk of bias

ROBINS-E: Risk of Bias in Non-randomized Studies of Exposures

ROBINS-I: Risk of Bias in Non-randomized Studies of Interventions

SoF: Summary of Findings

SR: systematic review

SYRCLE: SYstematic Review Center for Laboratory animal Experimentation

T4: thyroid hormone thyroxine

TSH: thyroid simulation hormones

uBPA: bisphenol-A level measured in urinary output

WHO: World Health Organization

DECLARATION OF ACADEMIC ACHIEVEMENT

I declare that I, jointly with my supervisor, Professor Holger J. Schünemann, played the primary role in the conception, design, and execution of the studies here included. We obtained feedback and advice from Professors Santesso and Holloway, and Dr. Thayer, as well as from members of the GRADE Working Group.

This work is original research that I conducted. I am the principle contributor and first author of all the manuscripts contained in this dissertation.

I am responsible and made the following contributions in all projects included in this work: design, conception, analysis, and writing of materials; I designed the guidance and data extraction for recording judgments from raters, and synthesized results for the risk of bias instrument development. I performed the reliability and validity analyses, and developed the surveys for the comparison across risk of bias instruments. I reviewed comments and feedback from experts during meetings with topic-specific experts in the environmental health field and the GRADE Working Group.

I conducted all analyses, designed figures and tables, and organized meetings. I wrote the manuscripts with editorial advice and supervision of Professor Schünemann, and with feedback from Professors Santesso and Holloway, and Dr. Thayer. The authors on each paper contributed significantly with important comments and advice for the final manuscripts.

For all the four manuscripts composing this “sandwich” thesis, earlier drafts of parts of this research have been presented at international academic conferences as part of the

manuscripts' development. The first paper was published in Environment International in 2016. The second paper was submitted to Environment International in December 2017 and is under review. The third paper is the final draft of the official GRADE guidance, which will be submitted to the GRADE Working Group for review and approval in March of 2018. The fourth paper will be submitted to Environment International.

CHAPTER 1. INTRODUCTION

1. Advancing systematic-review and decision-making methods in environmental health

"There is high demand in environmental and occupational health for using systematic review methodology and structured frameworks to evaluate and integrate evidence to support evidence-based and transparent decisions and recommendations" [1]. In 2016, I postulated that the release of a risk of bias (RoB) instrument that evaluated the RoB of non-randomized studies (NRS) against an ideal target trial could change the methodological approach when making decisions about environmental exposures. While gaining traction for use in the environmental health field, there was hesitation among environmental health researchers to use established decision-making frameworks, such as GRADE (Grading of Recommendations Assessment, Development and Evaluation). Some concerns stemmed from a lack of familiarity with the framework, limited exploration into the integration of current RoB instruments into GRADE, and discomfort from considering well-conducted NRS of exposure at 'Low' certainty of evidence (CoE) within GRADE due to bias resulting from prognostic imbalance and confounding. This instrument, the recently released RoB in NRS of Interventions (ROBINS-I), had undergone careful development and piloting on NRS of health interventions [2]. A key feature of ROBINS-I is establishing a structured comparison of RoB by evaluating the RoB of NRS studies against an ideal target trial as a reference for low risk of bias. Since the domains within ROBINS-I overlapped with concepts in other instruments commonly used to evaluate studies of environmental exposure, I expected that ROBINS-I could be applied to studies of exposures; however, I did not know if the

instrument could be applied verbatim or whether modifications might be needed to improve understanding.

Indeed, adaptation of terminology for exposure studies may be desirable [3-7]. For example, in the ROBINS-I instrument, the term “intervention” is used to refer to “treatment” or “exposure” groups in NRS. “Exposure” in the case of the ROBINS-I instrument represents a voluntary medical intervention or treatment (i.e., prenatal folic acid supplementation), not an unintentional exposure to an environmental or occupational hazard. In addition, fundamental challenges with the evaluation of unintentional exposures includes limited information on the start and duration of the exposure, certainty in the measurement of the exposure, and distinction between different levels of exposure. For example, bisphenol A (BPA), a chemical commonly used to make polycarbonate plastics, is considered ubiquitous in the environment and widespread among humans. Therefore, studies evaluating the effect of BPA on health outcomes may be able to detect current levels of BPA in the body, but may not be able to determine when exposure to BPA started or address how the exposure of each individual differs over time.

Multiple instruments have been used to assess the risk of bias, often called study quality or internal validity, in NRS [8, 9] and, due to a lack of clear advantages of one instrument over another, no single instrument is strongly recommended for use in systematic review [10]. This lack of guidance on what instrument to use is a key issue in environmental health, where NRS predominate. Rooney et al. examined risk of bias instruments and criteria used by five different organizations to evaluate RoB in NRS of

exposures: the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) Working Group; the National Toxicology Program's (NTP) Office of Health Assessment and Translation (OHAT); the UCSF Navigation Guide; the NTP's Office of the Report on Carcinogens (ORoC); and the Integrated Risk Information System of the U.S. Environmental Protection Agency (EPA-IRIS) [11]. The authors identified many similarities between instruments at the domain and question level, such as tailoring of the instrument to the design of the study, and merging with GRADE's suggestion to assess risk of bias on an outcome level was noted. Also, there appeared to a merging of ideas in the domains of assessment (e.g., participant selection, confounding, attrition/exclusion, exposure/intervention assessment, outcome assessment and selective reporting). However, the authors noted a few differences between the instruments, namely within the specific items used for risk of bias assessment, whether or not to reach overall study ratings, and the procedures for evaluation of risk of bias across studies in a systematic review.

Suggesting a single instrument is not without challenges [11]. Surmountable barriers include ensuring a common understanding of terminology and definitions, evaluating study limitations or strengths that encompass more than one domain, and characterizing complex issues such as impact of confounding or quality of exposure assessment within a structured approach. In addition to these logical arguments, empirical evidence, e.g. about the reliability and validity of the instruments, that allows for comparing different instruments would support expressing a preference of one over another instrument. However, little evidence exists of reliability or validity testing for

any RoB instrument that addresses exposure studies. Systematic-review authors and guideline developers would benefit from such empirical data to establish the importance of individual domains or distinguish between the performance of instruments.

RoB instruments play an important role in the evaluation of evidence to inform decisions. When using evidence from non-randomized studies (NRS) to answer questions about the effects of environmental exposures on health, it is important to assess RoB of individual studies as part of determining the certainty in the body of evidence. Using the fundamental design of ROBINS-I, I explored development of a version of the instrument to evaluate RoB in studies of environmental exposure. I identified important modifications necessitating a distinct instrument: The RoB instrument for NRS of exposures. Using the ideal target trial or target experiment to assess RoB with the RoB instrument for NRS of exposures has impact on how structured evidence synthesis frameworks, such as GRADE, will evaluate NRS.

2. Harmonization of systematic-review and decision-making methods

Efforts are ongoing to harmonize methods, many of which have fed back into this research. As mentioned in our first publication, "[i]n 2014, several project groups were formed within the GRADE Working Group to focus on methods assessment needs that are directly applicable to environmental and occupational health, including project groups for environmental health, observational studies, public health, application of

GRADE to laboratory animal research, and non-randomized study risk of bias integration." Methods advancements from those project groups include guidance for the application and integration of ROBINS-I within GRADE [12]; developments in preclinical animal intervention studies in the context of therapeutic interventions (animal group) [13]; instruments to facilitate the presentation of multiple evidence streams (i.e., human, animal, *in vitro*, and *in silico*) within GRADE's official Guideline Development Tool software GRADEpro (www.gradepro.org); and considerations for the integration of randomized and non-randomized study designs within systematic reviews and guidelines [14, 15].

I prioritized research at the intersection of methods development and environmental-health topic-specific expertise. During the pursuit of this project, I presented in multiple formats to members of the GRADE Working Group; attendees at international conferences, such as the Cochrane Collaboration, Guideline International Network Conference, and Environmental Protection Agency Workshop on Chemical Risk Assessment; and participants of the ROBINS for Exposure (ROBINS-E) work group. I solicited feedback and scenarios to enhance the accuracy and widen the applicability of our findings.

3. Goals and scope

This dissertation highlights the importance of standardized methods for environmental-health decision making in four stages:

1) I recognize the state of the science of evidence assessment and decision making in environmental health;
2) As a result of pilot testing and external feedback, I propose a modified instrument to evaluate the RoB of NRS of exposures: The RoB instrument for NRS of exposures;
3) To complement the development of a novel instrument for evaluation of RoB of NRS of exposure, I provide detailed guidance and examples for the implementation of the RoB instrument for NRS of exposures; and
4) To understand the reliability and validity of the RoB instrument for NRS of exposures, I conduct multiple analyses on the interrater reliability and inter-instrument reliability when compared with other commonly used instruments in the environmental field.

Our intention is to deliver a robust instrument that can guide systematic-review authors and guideline developers when evaluating RoB in NRS of exposure and integrating those results into a decision-making framework, such as GRADE. In addition, I aim to address some of the concerns expressed in the environmental-health community on use of GRADE by presenting an instrument that measures all studies of exposure along a standard comparison with RCTs.

4. Thesis overview

As mentioned previously, this dissertation is organized in four main research sections with a fifth section for concluding thoughts, pulling these themes back together at the

end of the document. Chapter 2 highlights the state of the research of systematic-review and guideline-development methods for the environmental health field. I identify areas for advancement in environmental-health decision making, specifically when using the GRADE framework. Chapter 3 introduces our instrument to evaluate RoB in NRS: The RoB instrument for NRS of exposures. I present the results from piloting necessitating a distinct instrument for exposures, and the modifications made to ROBINS-I. Chapter 4 elaborates on the RoB instrument for NRS of exposures. I describe the process for using the RoB instrument for NRS of exposures within the GRADE framework. I intersperse the guidance with examples from the application of the RoB instrument for NRS of exposures and demonstrate the integration of the RoB instrument for NRS of exposures within the GRADE framework. Chapter 4 examines the robustness of the RoB instrument for NRS of exposures, calculating the interrater reliability and construct validity. I perform the same calculations on three other instruments commonly used to assess RoB within NRS of exposures and present comparisons across all four instruments. Within our concluding remarks, in Chapter 5, I reflect on our progress in this discipline since the publication of our second chapter. In addition, I present some challenges of the work and areas for continued advancement.

5. References

1. Morgan RL, Thayer KA, Bero L, Bruce N, Falck-Ytter Y, Ghersi D, Guyatt G, Hooijmans C, Langendam M, Mandrioli D *et al*: **GRADE: Assessing the quality of evidence in environmental and occupational health**. *Environ Int* 2016, **92-93**:611-616.

2. Sterne JA, Hernan MA, Reeves BC, Savovic J, Berkman ND, Viswanathan M, Henry D, Altman DG, Ansari MT, Boutron I *et al*: **ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions**. *BMJ* 2016, **355**:i4919.

3. NRC (National Research Council): **Review of EPA's Integrated Risk Information System (IRIS) Process (**http://www.nap.edu/catalog.php?record_id=18764**) [accessed 1 January 2015]**. 2014.

4. Johnson PI, Sutton P, Atchley DS, Koustas E, Lam J, Sen S, Robinson KA, Axelrad DA, Woodruff TJ: **The Navigation Guide - evidence-based medicine meets environmental health: systematic review of human evidence for PFOA effects on fetal growth**. *Environ Health Perspect* 2014, **122**(10):1028-1039.

5. Lam J, Koustas E, Sutton P, Johnson PI, Atchley DS, Sen S, Robinson KA, Axelrad DA, Woodruff TJ: **The Navigation Guide—evidence-based medicine meets environmental health: integration of animal and human evidence for PFOA effects on fetal growth**. *Environ Health Perspect* 2014, **122**(10):1040-1051.

6. Koustas E, Lam J, Sutton P, Johnson PI, Atchley DS, Sen S, Robinson KA, Axelrad DA, Woodruff TJ: **The Navigation Guide - evidence-based medicine meets environmental health: systematic review of nonhuman evidence for PFOA effects on fetal growth**. *Environ Health Perspect* 2014, **122**(10):1015-1027.

7. NTP (National Toxicology Program): **Handbook for Conducting a Literature-Based Health Assessment Using Office of Health Assessment and Translation (OHAT) Approach for Systematic Review and Evidence Integration. January 9, 2015 release. Available at** http://ntp.niehs.nih.gov/go/38673**.** 2015.

8. Shamliyan T, Kane RL, Dickinson S: **A systematic review of tools used to assess the quality of observational studies that examine incidence or prevalence and risk factors for diseases**. *J Clin Epidemiol* 2010, **63**(10):1061-1070.

9. Deeks JJ, Dinnes J, D'amico R, Sowden A, Sakarovitch C, Song F, Petticrew M, Altman D: **Evaluating non-randomised intervention studies**. *Health technology assessment* 2003, **7**(27):1-179.

10. Higgins J, Green S: **Cochrane Handbook for Systematic Reviews of Interventions. Version 5.1.0 (updated March 2011).** http://handbook.cochrane.org/ **[accessed 3 February 2013]**. 2011.

11. Rooney AA, Cooper GS, Jahnke GD, Lam J, Morgan RL, Boyles AL, Ratcliffe JM, Kraft AD, Schünemann HJ, Schwingl P: **How credible are the study results? Evaluating and applying internal validity tools to literature-based assessments of environmental health hazards**. *Environment international* 2016.

12. Schünemann H, Cuello C, Akl EA, Mustafa R, Meerpohl J, Thayer K, Morgan R, Gartlehner G, Kunz R, Katikireddi S *et al*: **GRADE Guidelines: 18. How tools to assess risk of bias in non-randomized studies should be used to rate the certainty of a body of evidence**. Unpublished.

13. Hooijmans C, DeVries R, Ritskes-Hoitinga M, Rovers M, Leeflang MM, Inthout J, Wever K, Hooft L, Jadebeer JK, T., Macleod M *et al*: **Facilitating healthcare decisions by assessing the certainty in the evidence from preclinical animal studies**. *PLOS ONE* Accepted for publication.

14. Cuello C, Morgan RL, Guyatt G, Brozek J, Santesso N, Thayer K, Verbeek JH, Schunemann HJ: **A scoping review and survey provides the rationale, perceptions, and preferences for the integration of randomized and non-randomized studies in evidence syntheses**. *Journal of clinical epidemiology* Under review.

15. Cuello C, Morgan RL, Verbeek JH, Guyatt G, Ansari MT, Brozek J, Thayer K, Schünemann HJ: **Strategies to optimize use of randomized and non-randomized studies in evidence syntheses that use GRADE**. *Journal of clinical epidemiology* Under review.

CHAPTER 2. GRADE: ASSESSING THE QUALITY OF EVIDENCE IN ENVIRONMENTAL AND OCCUPATIONAL HEALTH

PREFACE TO CHAPTER 2

Chapter 2. *GRADE: Assessing the quality of evidence in environmental and occupational health* was submitted to Environment International on 30 July 2015, submitted in revised form on 24 November 2015, and accepted for print on 10 January 2016. The final manuscript was available online on 27 January 2016. In this dissertation, we present the revised submitted version.

GRADE: Assessing the quality of evidence in environmental and occupational health

Authors

Rebecca L. Morgan [a]; Kristina A. Thayer [b]; Lisa Bero [c]; Nigel Bruce [d]; Yngve Falck-Ytter [e]; Davina Ghersi [f], Gordon Guyatt [a]; Carlijn Hooijmans [g]; Miranda Langendam [h]; Daniele Mandrioli [i]; Reem A. Mustafa [a,j]; Eva A. Rehfuess [k]; Andrew A. Rooney [b]; Beverley Shea [l]; Ellen K. Silbergeld [m]; Patrice Sutton [n]; Mary S. Wolfe [b]; Tracey J. Woodruff [n]; Jos H Verbeek [o]; Alison C. Holloway [p]; Nancy Santesso [a]; Holger J. Schünemann [a,q]

Order of Authors: Rebecca L Morgan, MPH; Kristina A Thayer , PhD; Lisa Bero , PhD; Nigel Bruce, MBChB, PhD; Yngve Falck-Ytter, MD; Davina Ghersi, MPH, PhD; Gordon Guyatt, MD, MSc; Carlijn Hooijmans, PhD; Miranda Langendam, PhD; Daniele Mandrioli, MD; Reem A. Mustafa, MD, MPH, PhD; Eva A Rehfuess, MA (Oxon), PhD; Andrew A Rooney, PhD; Beverley Shea, PhD; Ellen K Silbergeld, PhD; Patrice Sutton, MPH; Mary Wolfe, PhD; Tracey J Woodruff, MPH, PhD; Jos H Verbeek, MD, PhD; Alison C. Holloway, PhD; Nancy Santesso, RD, PhD; Holger J Schünemann, MD, PhD, M.Sc.

Author Affiliations

[a] Department of Clinical Epidemiology & Biostatistics, McMaster University, Health Sciences Centre, Room 2C14, 1280 Main Street West, Hamilton, ON L8S 4K1 Canada

morganrl@mcmaster.ca, guyatt@mcmaster.ca, santesna@mcmaster.ca, schuneha@mcmaster.ca

[b] Division of the National Toxicology Program, National Institute of Environmental Health Sciences, National Institutes of Health, Department of Health and Human Services, P.O. Box 12233, Mail Drop K2-02, Research Triangle Park, NC USA 27709. thayer@niehs.nih.gov, andrew.rooney@nih.gov, wolfe@niehs.nih.gov

[c] Charles Perkins Centre, The University of Sydney, D17, The Hub, 6th floor, Charles Perkins Centre, The University of Sydney, New South Wales, 2006. lisa.bero@sydney.edu.au

[d] Department of Public Health and Policy, University of Liverpool, L69 3GB, United Kingdom. ngb@liv.ac.uk

[e] Division of Gastroenterology, Case Western Reserve University and Louis Stokes VA Medical Center, Cleveland, 10701 East Blvd., Cleveland, OH, 44106, USA. Yngve.Falck-Ytter@case.edu

[f] Sydney Medical School, University of Sydney, NSW 2006; National Health and Medical Research Council, 16 Marcus Clarke Street, Canberra City, ACT 2601, Australia. Email davina.ghersi@nhmrc.gov.au

[g] Departments of SYRCLE and Anesthesiology, Radboud University Medical Centre, Geert Grooteplein-Noord 29, route 231, 6525 GA Nijmegen, The Netherlands. Carlijn.Hooijmans@radboudumc.nl

[h] Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, University of Amsterdam, Room J1B-211, P.O. Box 22660, 1100 DD Amsterdam, The Netherlands. m.w.langendam@amc.uva.nl

[l] Cesare Maltoni Cancer Research Center, Ramazzini Institute, Via Saliceto 3, Bentivoglio (Bologna), P.O. Box 40133, Italy. mandriolid@ramazzini.it

[j] Departments of Medicine/Nephrology and Biomedical & Health Informatics, University of Missouri-Kansas City, School of Medicine, M4-303, 2411 Holmes St., Kansas City, Missouri 64108-2792. ramustafa@gmail.com

[k] Institute for Medical Informatics, Biometry and Epidemiology, University of Munich, Marchioninistr. 15, 81377 Munich, Germany. rehfuess@ibe.med.uni-muenchen.de

[l] Bruyere Research Institute and Ottawa Hospital Research Institute, University of Ottawa, Ottawa, ON. bevshea@uottawa.ca

[m] Department of Environmental Health Sciences, Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe Street, E6644, Baltimore, Maryland 21205, USA. esilber2@jhu.edu

[n] Program on Reproductive Health and the Environment, University of California-San Francisco, 550 16th Street, San Francisco, California USA, 94143. patrice.sutton@ucsf.edu, tracey.woodrfuff@ucsf.edu

[o] Finnish Institute of Occupational Health, Cochrane Work, PO Box 310, 70101 Kuopio, Finland. Jos.Verbeek@ttl.fi

[p] Department of Obstetrics and Gynecology, McMaster University, Health Sciences Centre, Room 3N52A, 1280 Main Street West, Hamilton, ON L8S 4K1 Canada. hollow@mcmaster.ca

[q] Department of Medicine, McMaster University, Health Sciences Centre, Room 2C14, 1280 Main Street West, Hamilton, ON L8S 4K1 Canada.

holger.schunemann@mcmaster.ca

Corresponding author

Holger J. Schünemann. Department of Clinical Epidemiology & Biostatistics, Health Sciences Centre, Room 2C14, 1280 Main Street West, Hamilton, ON L8S 4K1 Canada.

holger.schunemann@mcmaster.ca

Conflict of interest: The authors declare they have no financial interests with respect to this manuscript, or its content, or subject matter.

Abstract

There is high demand in environmental health for adoption of a structured process that evaluates and integrates evidence while making decisions and recommendations transparent. The Grading of Recommendations Assessment, Development and Evaluation (GRADE) framework holds promise to address this demand. For over a decade, GRADE has been applied successfully to areas of clinical medicine, public health, and health policy, but experience with GRADE in environmental and occupational health is just beginning. Environmental and occupational health questions focus on understanding whether an exposure is a potential health hazard or risk, assessing the exposure to understand the extent and magnitude of risk, and exploring interventions to mitigate exposure or risk. Although GRADE offers many advantages, including its flexibility and methodological rigor, there are features of the different sources of evidence used in environmental and occupational health that will require further consideration to assess the need for method refinement. An issue that requires particular attention is the evaluation and integration of evidence from human, animal, *in vitro,* and *in silico* (computer modelling) studies when determining whether an environmental factor represents a potential health hazard or risk. Assessment of the hazard of exposures can produce analyses for use in the GRADE evidence-to-decision (EtD) framework to inform risk-management decisions about removing harmful exposures or mitigating risks. The EtD framework allows for grading the strength of the recommendations based on judgments of the certainty in the evidence (also known as quality of the evidence), as well as other factors that inform recommendations such as

social values and preferences, resource implications, and benefits. GRADE represents an untapped opportunity for environmental and occupational health to make evidence-based recommendations in a systematic and transparent manner. The objectives of this article are to provide an overview of GRADE, discuss GRADE's applicability to environmental health, and identify priority areas for method assessment and development.

Keywords: GRADE; Evidence-based; Risk of Bias; Environmental Health; Risk Assessment; Recommendations

Highlights

- A structured framework is needed for decision-making in environmental health.
- GRADE has been applied in many disciplines and holds great promise for the field.
- Methods development and assessment is needed to address environmental health data.
- Methods assessment priorities are evaluation and integration of diverse evidence streams.
- GRADE evidence-to-decision framework informs risk and other management decisions.

Abbreviations

AHRQ - Agency for Healthcare Research and Quality

ASTDR – Agency for Toxic Substances and Disease Registry

CDC – Centers for Disease Control and Prevention

CiE – Certainty in the Evidence

EFSA - European Food Safety Authority

EPA - Environmental Protection Agency

EtD – Evidence-to-decision

GRADE - Grading of Recommendations Assessment, Development, and Evaluation working group

OHAT - Office of Health Assessment and Translation

PECO – Population, Exposure, Comparator, Outcome

PICO – Population, Intervention, Comparator, Outcome

NRC – National Research Council

NTP - National Toxicology Program

RoB – Risk of Bias

SYRCLE - SYstematic Review Center for Laboratory animal Experimentation

WHO – World Health Organization

1 Introduction

There is high demand in environmental and occupational health for using systematic review methodology and structured frameworks to evaluate and integrate evidence to support evidence-based and transparent decisions and recommendations [Agency for Toxic Substances and Disease Registry (ATSDR) 1, 2, 3, NRC 4, 5, EFSA 6, 7-16]. Environmental health, which includes occupational health, is a broad field in which data address all the physical, chemical, and biological factors external to a person, and all the related factors impacting behaviors [17]. Environmental health questions focus on understanding whether an exposure is a potential health hazard or risk using exposure assessments to recognize the extent and magnitude of exposure, and interventions to prevent or mitigate exposure or risk.

The Grading of Recommendations Assessment, Development, and Evaluation (GRADE) approach has the potential to improve transparency in addressing these questions in environmental health assessments. GRADE represents a rigorous, structured, and transparent process to inform decision-making beginning with well-defined questions, followed by an assessment of the certainty in the evidence (also called confidence in the effect or other estimates, or quality of the evidence) [18, 19], and leading to development of recommendations and decisions.

GRADE is widely used internationally to address topics related to clinical medicine, public health, and health policy [19-22], including by programs within the U.S. Centers for Disease Control and Prevention (CDC), World Health Organization (WHO), the U.S. Agency for Healthcare Research and Quality (AHRQ), and National Institute for Health

and Clinical Excellence (NICE) in the United Kingdom and the National Health and Medical Research Council in Australia [23-27]. The Cochrane Collaboration, which prepares, maintains, and promotes the accessibility of systematic reviews, uses the GRADE system for reporting on the quality of evidence for outcomes in systematic reviews [28, 29]. Formed in 2000, the GRADE Working Group now includes over 500 active members from 40 countries and serves as a think tank for advancing evidence-based decision-making in multiple disciplines [18](see also http://www.gradeworkinggroup.org/).

Advantages of using the GRADE approach have already been recognized by some within the environmental health field. The Navigation Guide proposed adapting GRADE for an environmental health context [7] and followed-up with a series of case studies to demonstrate the feasibility of applying GRADE to epidemiological and animal studies [11-13, 30]. In 2013, the National Toxicology Program's (NTP) Office of Health Assessment and Translation (OHAT) at the National Institute of Environmental Health Sciences announced plans to use GRADE in its evaluations to assess the evidence for associations between environmental exposures and non-cancer health effects [NTP 31, 32, NTP 33]. The SYstematic Review Center for Laboratory animal Experimentation (SYRCLE), is currently applying the GRADE approach to assess the quality of evidence from preclinical animal intervention studies [34]. GRADE has also been used in recent systematic reviews of epidemiological studies of shift work and breast cancer risk [35], shift work and cardiovascular disease [36], and adverse effects related to reduced indoor air quality related to household fuel use [37, 38]. GRADE, including its adoption

by NTP/OHAT and the Navigation Guide, was specifically identified in the National Academy of Sciences' National Research Council (NRC) review of the U.S. Environmental Protection Agency's (EPA) Integrated Risk Information System as an approach that would increase the transparency of evaluating evidence [14]. Use of GRADE in environmental health is likely to grow as systematic reviews become more common in the field and the limitations of expert-based narrative review methods are increasingly recognized [4, 6, 8, 10, 39, 40].

An additional advantage of GRADE is the GRADE Working Group's commitment to ongoing methods development and assessment of applicability to different areas of research. This is critical because experience with GRADE in the environmental health context is limited. Work to-date from the Navigation Guide, NTP, and WHO show the GRADE framework is sufficiently flexible to support use now [11-13, 33, 37, 41]; however, areas for further method assessment have been identified. In this respect, the GRADE Working Group serves as a vehicle to leverage transdisciplinary skills, knowledge, and resources to bridge the fields of clinical and environmental health. The objectives of this article are to provide an overview of the GRADE framework, discuss applicability of GRADE to environmental and occupational health, and identify priority areas for method development.

2 GRADE Approach

2.1 Formulating the Research Question

GRADE requires that decision-makers specify key-elements to formulate a relevant and focused question for decision-making (e.g., to inform clinical and public health guidelines, formulate scientific consensus statements, etc.) [39, 42]. The key elements are the components of the question that identify what information must be provided in a primary study to evaluate the intervention under assessment and hence answer the question [39]. For instance, for questions aimed at evaluating interventions, the key elements are the Population, Intervention, Comparator, and Outcome (PICO) [42, 43]. Both beneficial and harmful outcomes that the target population may experience as a result of the intervention should be considered. At present, GRADE focuses on answering decision-making (i.e., actionable) questions about interventions (including diagnostic tests and strategies), though the GRADE framework has been expanded to prognostic questions [44, 45].

2.2 Quality of the Evidence

GRADE uses a structured framework to determine overall certainty in the evidence (CiE) for outcomes across a collection of research studies or body of evidence (Figure 1)[46]. The GRADE approach does not remove judgment from decision-making; however, the approach provides a framework of critical components to assess, guidance on the consideration of empirical evidence, and emphasizes transparency throughout the process. An initial evaluation of the CiE is conducted based on whether or not the research studies used randomized allocation. In the current GRADE approach, the CiE from randomized controlled trials (RCT) receives an initial rating of “high”, whereas the CiE from observational (i.e., non-randomized) studies starts at “low”. After this initial

evaluation of randomization, other aspects of risk of bias (RoB), i.e., internal validity, are assessed. GRADE does not recommend the use of a specific RoB tool, but suggests specific criteria that should be considered when assessing a body of randomized or non-randomized studies that address risk of bias [47]. In addition to RoB, the certainty in a body of evidence can be rated down for inconsistency, indirectness, imprecision, or publication bias, or rated up for the magnitude of the effect, dose-response gradient, or direction and impact of residual plausible confounding. Different terminology may be used to describe these elements as long as the concepts are identical [46, 48]. Like RCTs, randomized experimental studies in animals would start as "high" and typically be downgraded for indirectness due to differences in the population [49]. The evidence is assessed and presented in an evidence summary table separately for each critical or important outcome and expressed using four levels of certainty ratings (i.e., "high", "moderate", "low", or "very low") [50, 51]. This table, called a GRADE Evidence Profile or Summary of Findings table, requires transparent descriptions of the reasons for rating down and rating up [37].

2.3 Recommendations and the Evidence-to-Decision Framework

In addition to assessing the CiE across outcomes, the GRADE EtD framework explicitly considers the balance of benefits and harms, values and preferences, resource implications, feasibility, equity, and acceptability to determine the strength of the recommendation (strong or weak), and the direction (for or against) to make a final recommendation or decision [52-54]. The elements of the framework's structure transparently display the important criteria for deliberation (including relevant research

evidence, judgments from decision makers, and other considerations) to inform the balance about the desirable and undesirable consequences of the options or interventions considered. A judgment is needed for making decisions during all steps. However, the GRADE EtD framework provides a structure to maximize transparency and limit subjectivity throughout the process: in fact CiE is a key determinant for making a strong GRADE guidelines recommendation [55].

3 Considerations for Environmental Health

3.1 Formulating the Research Question

The GRADE approach has been utilized predominantly to answer questions on interventions in health care, like "what is the impact of an intervention (including diagnostic tests and strategies) compared with an alternative on patient or population important outcomes?" or "should intervention A or B be used for X?" In the context of decision-making in environmental health, the term intervention has somewhat different connotations. First, an intervention can be thought of as a specific environmental factor (i.e., exposure) that is being evaluated in human, animal, *in vitro*, or *in silico* studies as a risk factor or causative agent for an undesirable health outcome. In this scenario, the PICO question can be rephrased as a PECO question, where the term "Intervention" is replaced with "Exposure" [8, 33, 56]. The complexity of the exposure questions will vary, ranging from a single well-defined chemical to complex scenarios like wind farms, agricultural run-off, etc. To address the benefits and harms to humans from wind farms, PECO questions were developed to look at the exposure of physical emissions produced by wind farms or wind turbines (e.g., noise, infrasound, shadow flicker, and

electromagnetic radiation), as compared with no exposure to the physical emissions produced by wind farms or turbines [57]. Questions assessing exposures as risk factors or causative agents are used in risk assessments, which have several sub-questions [58, 59]:

- Hazard identification: What health problems are caused by the environmental factor?
- Dose-response assessment: What are the health problems at different exposure levels?
- Exposure assessment: What is the extent and nature of the exposure in the target population?
- Risk characterization: What is the extra risk of health problems in the exposed population?

Second, an environmental intervention question could be formulated to evaluate the impact of interventions that prevent or mitigate an exposure or risk. Environmental exposure-related interventions typically address chemical or physical agents in the environment, such as air, soil, water, or food, in a public or occupational setting, with the goal of trying to prevent, remove, or reduce exposure levels (e.g., reduction at source, improved ventilation, ingredient reformulation) through regulatory, technical, or behavioral interventions. Questions assessing the effects of an intervention to prevent or reduce exposure should be based on an established relationship between the exposure and health outcome(s). For example, since the relationship between noise exposure and noise-induced hearing loss has been established, showing that an

intervention reduces noise exposure is sufficient to also to conclude that the intervention decreases noise-induced hearing loss [60]. In studies of environmental health, such questions have the ability to compare the desirable consequences of reducing an exposure with potentially undesirable consequences of removing an exposure (e.g., costs, use of alternatives with unknown toxicity). While these types of questions are very similar to the clinical or public health intervention PICO questions GRADE was designed to assess, some challenges have been identified, such as how to assess complex interventions, use non-epidemiological evidence, and choosing outcomes and outcome measures [61]. Methodological research has continued to address concerns with applying GRADE to studies of interventions [42, 62].

3.2 Quality of the Evidence

3.2.1 Human and Experimental Animal Data

In environmental health, observational human studies and experimental animal studies (where animals are randomly assigned to treatment groups), and observational animal studies (i.e., "wildlife studies" or natural population-based studies) are often the highest quality evidence available to understand *whether* there is an association (or, if possible, cause-effect relationship) between an exposure and health outcome, as in the case of carcinogens [63]. The factors considered in GRADE when making and presenting judgments about the CiE (Figure 1) translate well to observational human and experimental animal studies, although harmonization of RoB tools and development of additional guidance on when rating down or rating up should be pursued. The WHO considered evidence from both non-randomized experimental and observational studies

to inform their Recommendations for Indoor Air Quality [37]. In the report, WHO assessed whether or not coal should be used as a household fuel. The decision to recommend against using unprocessed coal as a household fuel was informed by 1) the results from studies of cancer in humans and experimental animals; 2) systematic reviews of observational studies on particulate matter exposure and risk of lung cancer; and 3) population-level studies on the toxicity of coal and the impact of banning coal. While possible confounders of the different study types were recognized, they still provided the best available evidence to inform the recommendations. In addition, on-going methods development for rating the risk of bias [11-13, 33, 37, 64, 65] includes searching for observational studies that might be considered equivalent to randomized trials for the initial assessment of the risk of bias (e.g., factors in study design and execution that mitigate the lack of randomization, such as steps taken to fully control or adjust for confounding). Examples, however are currently lacking.

3.2.2 Mechanistic Data

In environmental health, human and experimental animal data are often interpreted in conjunction with evidence from mechanistic data supporting the biological plausibility of an association and/or to prioritize chemicals for additional testing or evaluation. The GRADE framework does not explicitly address mechanistic data, but they may be used to inform judgments about indirectness. There are an estimated 85,000 chemicals in commerce, the vast majority of which have not been tested for toxicity, even though in many cases the evidence available for a chemical will be mechanistic in nature [66, 67]. The lack of toxicity data for most environmental chemicals has led to major initiatives to

generate high throughput screening (HTS) data for chemicals. For example, the NTP's Tox21 HTS program has generated data for ~10,000 chemicals on ~75 biochemical- and cell-based assays that cover a range of activities including overall cellular health (cytotoxicity and apoptosis induction, mitochondrial toxicity, DNA damage), perturbation of cell signaling pathways, inflammatory response induction, agonists/antagonists for 15 nuclear receptors, and drug metabolism [68]. The US EPA's ToxCast HTS program currently has mechanistic data on 1860 chemicals tested in up to 821 assay endpoints [69]; however, many chemicals are still untested. Computer-modeling approaches are also being pursued to predict potential hazard and likelihood of significant exposure. For mechanistic data, tools to rate RoB for *in vitro* and *in silico* studies need to be developed and their contribution to the stream of evidence for different outcomes should be determined because these data are expected to be used more widely for prioritizing chemicals of concern as well as replacing traditional data in regulatory assessments [10, 15]. When assessing the effects of wind farms on human health, both direct and indirect evidence was considered to address the PECO question [57]. When assessing the body of evidence across the outcome of shadow flicker, there was low quality direct evidence available; however, available indirect data suggested that shadow flicker can affect health by inducing seizures among persons prone to photosensitive epilepsy. The utility of the GRADE rating down and rating up factors also needs to be assessed, although the concepts should generally apply (e.g., magnitude of effect can be analogous to efficacy and potency in an *in vitro* system). Analyses to assess the predictive utility of mechanistic data are a high priority in toxicology, and results will inform indirectness ratings within the GRADE framework.

3.3 Evidence-to-Decision Frameworks

Very little work has been done to use structured and transparent decision-making frameworks to guide the development of recommendations in environmental health. The WHO Recommendations for Indoor Air Quality applied the GRADE EtD framework to guide their final recommendations [37]. For their recommendation on household use of coal, in addition to the quality of evidence from studies on carcinogenicity of coal, risk of lung cancer, and population-level studies on toxicity, they also determined that the benefits of replacing unprocessed coal with cleaner alternatives clearly outweigh the harms of replacement, the values and preferences of replacing coal varied among stakeholders, and that there may be some limitations to the feasibility of implementing cleaner alternatives based on affordability and supply. The GRADE EtD framework, which has the capacity to integrate consideration of the CiE of a health hazard with evidence of benefit associated with mitigating exposure, values, preferences, resource implications and other criteria, has great potential for enhancing the transparency of decision-making in environmental and occupational health. The strength of the recommendation may be apparent and actionable, or application of GRADE may reveal gaps in our knowledge, and thus help efficiently and effectively target the allocation of scarce research funds.

The regulation of diesel is an example of an environmental topic that could be addressed with the GRADE EtD framework. Diesel engine exhaust is carcinogenic to humans and associated with increased hospital admissions, emergency room visits, asthma attacks, and premature death [70, 71]. At the same time, diesel engines have

desirable consequences of higher fuel efficiency, lower carbon dioxide emissions, heavy duty hauling capacity, and durability. For example, EPA rule-making for diesel standards included consideration of the composition of diesel, technological feasibility, costs of retrofitting or replacing, cost-benefit analyses that include quantifying human health impacts, overall economic impact and alternatives assessment. Moreover, the rule-making applied to specific scenarios such as vehicles on highways, city streets, construction sites, and ports. These analyses have led to a number of emission standards for diesel fuel and diesel engines [72]. By 2030, EPA estimates that particulate matter and nitrous oxides will be reduced by 380,000 tons/year and 7 million tons/year, respectively. This will result in annual benefits of over $290 billion, at a cost of approximately $15 billion. The GRADE EtD framework could also be applied to alternative assessments that look for safer chemicals by identifying and evaluating the safety of alternative chemicals [73]. Although such assessments are often not regulatory, they are used to inform consumer choice and encourage industry to move to safer alternatives and can complement regulatory actions.

The challenges of applying the GRADE EtD framework to environmental health topics are expected to be similar to clinical research, with most findings requiring a careful weighing of the health and other benefits or harms. A challenge specific to decision-making for environmental health is that many regulatory agencies require a determination of an allowable level or threshold of an exposure or risk, while in other cases there is no allowable exposure (for example asbestos ban). In studies where there is not a clear desirable effect of the exposure, the balance may focus on how frequently

the undesirable effects occur. Research is also needed to increase understanding and acceptability of the format that desirable and undesirable consequences are presented in to end-users.

4 Future Directions

This paper provides an overview of important aspects of adapting GRADE to decision-making in environmental health. In 2014, several project groups were formed within the GRADE Working Group to focus on methods assessment needs that are directly applicable to environmental and occupational health, including project groups for environmental health, observational studies, public health, application of GRADE to laboratory animal research, and non-randomized study risk of bias integration. Priority areas for the environmental and occupational health project group include (1) developing approaches to evaluate and integrate evidence from observational human, animal, *in vitro*, and *in silico* (computer modeling) studies to determine whether an association exist between exposure and health outcome(s); (2) applying GRADE to evaluations of interventions to mitigate exposure or reduce risk when an association has been identified; and (3) gaining experience in applying the GRADE frameworks for evidence-to-decision (EtD) and determining the direction and strength of recommendations for environmental and occupational health topics. Critically adapting GRADE to environmental health also requires consideration of how to rate the overall strength of the evidence and to integrate evidence across multiple evidence streams.

5 Conclusions

This paper examines several key components of GRADE as they can be assessed and expanded as a standardized methodology for research and decision-making in environmental and occupational health. Over 90 organizations from 18 countries worldwide have adopted the GRADE framework to assess evidence and inform decision-making. With a focus on rigorous and transparent methods, the GRADE approach has been applied successfully to clinical medicine, public health, diagnostic decision-making, questions about prognosis, and has great potential for the field of environmental and occupational health. In parallel to the methods development that has occurred over the past decades in the clinical and public health field, environmental health scientists have developed topic specific expertise about the evidence that informs how the environment shapes our health and sets the stage for knowledge transfer across disciplines to strengthen the scientific basis of decision-making for public policy. Leveraging this synergy will increase the transparency of, and scientific basis for, decision-making in environmental health, and thus help secure improved health outcomes for individuals and populations.

6 Acknowledgments

This research was supported by the intramural research program of the National Institute of Environmental Health Sciences and the MacGRADE center at McMaster University. The contribution of UCSF Program on Reproductive Health and the Environment co-authors (TW and PS) to this research was supported by the Clarence Heller Foundation, the National Institute of Environmental Health Sciences (grants ES018135 and ESO22841), and U.S. EPA STAR grants (RD83467801 and RD83543301).

Authors would like to acknowledge the contributions of Elisa Aiassa and Annette

Martine Pruss-Ustun as members of the GRADE Environmental Health Project Group.

7 References

1. ATSDR: **The Future of Science at ATSDR: A Symposium**. In*: April 11 – 12, 2012 2012; Atlanta, GA*: US Department of Health and Human Services (DHHS) Agency for Toxic Substances and Disease Registry (ATSDR); 2012.
2. Mandrioli D, Sillbergeld E, Bero L: **Preperation of Evidence Based Toxicology Handbook.** https://colloquium.cochrane.org/meetings/evidence-based-toxicology-handbook**. Cochrane Colloquium expert meeting. Hyderabad, India (September 26, 2014)**. 2014.
3. Murray HE, Thayer KA: **Implementing systematic review in toxicological profiles: ATSDR and NIEHS/NTP collaboration**. *Journal of environmental health* 2014, **76**(8):34-35.
4. NRC: **Review of the Environmental Protection Agency's State-of-the-Science Evaluation of Nonmonotonic Dose-Response Relationships as they Apply to Endocrine Disrupters (**http://www.nap.edu/catalog.php?record_id=18608**) [accessed 1 January 2015]**. 2014.
5. Silbergeld E, Scherer RW: **Evidence-based toxicology: Strait is the gate, but the road is worth taking**. *Altex* 2013, **30**(1):67-73.
6. EFSA: **Application of systematic review methodology to food and feed safety assessments to support decision making**. *EFSA Journal* 2010, **8**(6):1637.
7. Woodruff TJ, Sutton P: **An evidence-based medicine methodology to bridge the gap between clinical and environmental health sciences**. *Health Affairs* 2011, **30**(5):931-937.
8. Woodruff TJ, Sutton P: **The Navigation Guide systematic review methodology: a rigorous and transparent method for translating environmental health science into better health outcomes**. *Environ Health Perspect* 2014, **122**(10):1007-1014.
9. Bruce N, Pope D, Rehfuess E, Balakrishnan K, Adair-Rohani H, Dora C: **WHO indoor air quality guidelines on household fuel combustion: Strategy implications of new evidence on interventions and exposure–risk functions**. *Atmospheric Environment* 2014, **Available online 27 August 2014**(0).
10. Mandrioli D, Silbergeld EK: **Evidence from Toxicology: The Most Essential Science for Prevention**. *Environmental health perspectives* 2015.
11. Johnson PI, Sutton P, Atchley DS, Koustas E, Lam J, Sen S, Robinson KA, Axelrad DA, Woodruff TJ: **The Navigation Guide - evidence-based medicine meets environmental health: systematic review of human evidence for PFOA effects on fetal growth**. *Environ Health Perspect* 2014, **122**(10):1028-1039.
12. Lam J, Koustas E, Sutton P, Johnson PI, Atchley DS, Sen S, Robinson KA, Axelrad DA, Woodruff TJ: **The Navigation Guide—evidence-based medicine meets environmental health: integration of animal and human evidence for PFOA effects on fetal growth**. *Environ Health Perspect* 2014, **122**(10):1040-1051.
13. Koustas E, Lam J, Sutton P, Johnson PI, Atchley DS, Sen S, Robinson KA, Axelrad DA, Woodruff TJ: **The Navigation Guide - evidence-based medicine meets**

environmental health: systematic review of nonhuman evidence for PFOA effects on fetal growth. *Environ Health Perspect* 2014, **122**(10):1015-1027.

14. NRC: **Review of EPA's Integrated Risk Information System (IRIS) Process (**http://www.nap.edu/catalog.php?record_id=18764**) [accessed 1 January 2015]**. 2014.
15. NRC: **Toxicity testing in the 21st century: A vision and a strategy**: National Academies Press; 2007.
16. Whaley P, Halsall C, Agerstrand M, Aiassa E, Benford D, Bilotta G, Coggon D, Collins C, Dempsey C, Duarte-Davidson R *et al*: **Implementing systematic review techniques in chemical risk assessment: Challenges, opportunities and recommendations**. *Environ Int* 2016, **92-93**:556-564.
17. **Environmental Health** [http://www.who.int/topics/environmental_health/en/]
18. Schünemann HJ, Best D, Vist G, Oxman AD: **Letters, numbers, symbols and words: how to communicate grades of evidence and recommendations**. *Canadian Medical Association Journal* 2003, **169**(7):677-680.
19. Guyatt GH, Oxman AD, Schunemann HJ, Tugwell P, Knottnerus A: **GRADE guidelines: A new series of articles in the Journal of Clinical Epidemiology**. *Journal of clinical epidemiology* 2011, **64**(4):380-382.
20. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, Schunemann HJ, Group GW: **GRADE: an emerging consensus on rating quality of evidence and strength of recommendations**. *BMJ* 2008, **336**(7650):924-926.
21. Schünemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Vist GE, Williams Jr JW, Kunz R, Craig J, Montori VM: **Grading quality of evidence and strength of recommendations for diagnostic tests and strategies**. *Bmj* 2008, **336**(7653):1106-1110.
22. Atkins D, Eccles M, Flottorp S, Guyatt GH, Henry D, Hill S, Liberati A, O'Connell D, Oxman AD, Phillips B: **Systems for grading the quality of evidence and the strength of recommendations I: critical appraisal of existing approaches The GRADE Working Group**. *BMC health services research* 2004, **4**(1):38.
23. National Health and Medical Research Council: **Procedures and requirements for meeting the 2011 NHMRC standard for clinical practice guidelines**. 2011.
24. Ahmed F, Temte JL, Campos-Outcalt D, Schünemann HJ, Group AEBRW: **Methods for developing evidence-based recommendations by the Advisory Committee on Immunization Practices (ACIP) of the US Centers for Disease Control and Prevention (CDC)**. *Vaccine* 2011, **29**(49):9171-9176.
25. Thornton J, Alderson P, Tan T, Turner C, Latchem S, Shaw E, Ruiz F, Reken S, Mugglestone MA, Hill J: **Introducing GRADE across the NICE clinical guideline program**. *Journal of clinical epidemiology* 2013, **66**(2):124-131.
26. Viswanathan M, Ansari MT, Berkman ND, Chang S, Hartling L, McPheeters M, Santaguida PL, Shamliyan T, Singh K, Tsertsvadze A: **Assessing the risk of bias of individual studies in systematic reviews of health care interventions**. 2012.
27. WHO: **WHO Handbook for guideline development**: World Health Organization; 2014.

28. Higgins JP, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, Savović J, Schulz KF, Weeks L, Sterne JA: **The Cochrane Collaboration's tool for assessing risk of bias in randomised trials**. *Bmj* 2011, **343**:d5928.
29. Schünemann H, Oxman A, Higgins J, Vist G, Glasziou P, Guyatt G: **Cochrane handbook for systematic reviews of interventions version 5.1. 0 (updated March 2011)**. 2011.
30. Vesterinen HM, Johnson PI, Atchley DS, Sutton P, Lam J, Zlatnik MG, Sen S, Woodruff TJ: **Fetal growth and maternal glomerular filtration rate: a systematic review**. *The Journal of Maternal-Fetal & Neonatal Medicine* 2014(0):1-6.
31. Program) NNT: **Board of Scientific Counselors June 25, 2013 meeting. Meeting materials available at** http://ntp.niehs.nih.gov/go/40246 **[accessed 9 August 2014]**. 2013.
32. Rooney AA, Boyles AL, Wolfe MS, Bucher JR, Thayer KA: **Systematic review and evidence integration for literature-based environmental health science assessments**. *Environ Health Perspect* 2014, **122**(7):711-718.
33. NTP (National Toxicology Program): **Handbook for Conducting a Literature-Based Health Assessment Using Office of Health Assessment and Translation (OHAT) Approach for Systematic Review and Evidence Integration. January 9, 2015 release. Available at** http://ntp.niehs.nih.gov/go/38673**.** 2015.
34. Hooijmans CR, Rovers MM, de Vries RB, Leenaars M, Ritskes-Hoitinga M, Langendam MW: **SYRCLE's risk of bias tool for animal studies**. *BMC medical research methodology* 2014, **14**(1):43.
35. Ijaz S, Verbeek J, Seidler A, Lindbohm M-L, Ojajarvi A, Orsini N, Costa G, Neuvonen K: **Night-shift work and breast cancer—a systematic review and meta-analysis**. *Scand J Work Environ Health* 2013, **39**(5):431-447.
36. Vyas MV, Garg AX, Iansavichus AV, Costella J, Donner A, Laugsand LE, Janszky I, Mrkobrada M, Parraga G, Hackam DG: **Shift work and vascular events: systematic review and meta-analysis**. *Bmj* 2012, **345**:e4800.
37. WHO: **Indoor air quality guidelines: household fuel combustion**; 2014.
38. Bruce N, Dora C, Krzyzanowski M, Adair-Rohani H, Morawska L, Wangchuk T: **Tackling the health burden from household air pollution: Development and implementation of new WHO Guidelines**. 2013.
39. Aiassa E, Higgins J, Frampton G, Greiner M, Afonso A, Amzal B, Deeks J, Dorne J-L, Glanville J, Lövei G: **Applicability and feasibility of systematic review for performing evidence-based risk assessment in food and feed safety**. *Critical reviews in food science and nutrition* 2015, **55**(7):1026-1034.
40. EPA: **Applying systematic review to assessments of health effects of chemical exposures**. In: *EPA Workshop: 2013; Washington, DC*; 2013.
41. Johnson PI, Sutton P, Atchley D, Koustas E, Lam J, Robinson K, Sen S, Axelrad D, Woodruff TJ: **Applying the Navigation Guide: Case Study #1: The impact of developmental exposure to perfluorooctanoic acid (PFOA) on fetal growth (Final protocol)** http://prhe.ucsf.edu/prhe/navigationguide.html **[accessed 29 November, 2014]** 2013.

42. Guyatt GH, Oxman AD, Kunz R, Atkins D, Brozek J, Vist G, Alderson P, Glasziou P, Falck-Ytter Y, Schunemann HJ: **GRADE guidelines: 2. Framing the question and deciding on important outcomes**. *J Clin Epidemiol* 2011, **64**(4):395-400.
43. Richardson WS, Wilson MC, Nishikawa J, Hayward RS: **The well-built clinical question: a key to evidence-based decisions**. *Acp j club* 1995, **123**(3):A12-13.
44. Spencer FA, Iorio A, You J, Murad MH, Schünemann HJ, Vandvik PO, Crowther MA, Pottie K, Lang ES, Meerpohl JJ: **Uncertainties in baseline risk estimates and confidence in treatment effects**. *Bmj* 2012, **345**:e7401.
45. Iorio A, Spencer FA, Falavigna M, Alba C, Lang E, Burnand B, McGinn T, Hayden J, Williams K, Shea B: **Use of GRADE for assessment of evidence about prognosis: rating confidence in estimates of event rates in broad categories of patients**. *bmj* 2015, **350**:h870.
46. Schünemann H, Brożek J, Oxman G: **Handbook for grading the quality of evidence and the strength of recommendations using the GRADE approach. 2013**. In.
47. Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, Montori V, Akl EA, Djulbegovic B, Falck-Ytter Y: **GRADE guidelines: 4. Rating the quality of evidence—study limitations (risk of bias)**. *Journal of clinical epidemiology* 2011, **64**(4):407-415.
48. GRADE Working Group: **Criteria for applying or using GRADE.** http://www.gradeworkinggroup.org/intro.htm#criteria **[accessed May 13, 2015]**. 2010.
49. Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, Alonso-Coello P, Falck-Ytter Y, Jaeschke R, Vist G *et al*: **GRADE guidelines: 8. Rating the quality of evidence--indirectness**. *Journal of clinical epidemiology* 2011, **64**(12):1303-1310.
50. Guyatt GH, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, Norris S, Falck-Ytter Y, Glasziou P, Debeer H *et al*: **GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables**. *Journal of Clinical Epidemiology* 2011, **64**(4):383-394.
51. Balshem H, Helfand M, Schunemann HJ, Oxman AD, Kunz R, Brozek J, Vist GE, Falck-Ytter Y, Meerpohl J, Norris S *et al*: **GRADE guidelines: 3. Rating the quality of evidence**. *J Clin Epidemiol* 2011, **64**(4):401-406.
52. Treweek S, Oxman AD, Alderson P, Bossuyt PM, Brandt L, Brozek J, Davoli M, Flottorp S, Harbour R, Hill S *et al*: **Developing and Evaluating Communication Strategies to Support Informed Decisions and Practice Based on Evidence (DECIDE): protocol and preliminary results**. *Implement Sci* 2013, **8**:6.
53. Schünemann HJ, Oxman AD, Vist GE, Higgins JPT, Deeks JJ, P. G, Guyatt GH, on behalf of the Cochrane Applicability and Recommendations Methods Group: **Chapter 12: Interpreting results and drawing conclusions**. In: *Cochrane Handbook for Systematic Reviews of Interventions Version 510 [updated March 2011].* Edited by Higgins JPT, Green S, eds.: The Cochrane Collaboration, 2011. Available at www.cochrane-handbook.org. [accessed 13 July 2012]; 2012.
54. Andrews J, Guyatt G, Oxman AD, Alderson P, Dahm P, Falck-Ytter Y, Nasser M, Meerpohl J, Post PN, Kunz R: **GRADE guidelines: 14. Going from evidence to**

recommendations: the significance and presentation of recommendations. *Journal of clinical epidemiology* 2013, **66**(7):719-725.

55. Djulbegovic B, Kumar A, Kaufman RM, Tobian A, Guyatt GH: **Quality of evidence is a key determinant for making a strong GRADE guidelines recommendation**. *Journal of clinical epidemiology* 2015.
56. Evidence. CfE: **Guidelines for Systematic Review and Evidence Synthesis in Environmental Management**. In: *Environmental Evidence.* vol. Version 4.2. www.environmentalevidence.org/Documents/Guidelines/Guidelines4.2.pdf: Collaboration for Environmental Evidence.; 2013.
57. Merlin T, Newton S, Ellery B, Milverton J, Farah C: **Systematic review of the human health effects of wind farms**. 2015.
58. **Hazard Identification** [http://www.epa.gov/risk_assessment/hazardous-identification.htm]
59. Schünemann H, Hill S, Guyatt G, Akl EA, Ahmed F: **The GRADE approach and Bradford Hill's criteria for causation**. *Journal of epidemiology and community health* 2011, **65**(5):392-395.
60. Verbeek JH, Kateman E, Morata TC, Dreschler WA, Mischke C: **Interventions to prevent occupational noise-induced hearing loss**. *The Cochrane Library* 2012.
61. Rehfuess EA, Akl EA: **Current experience with applying the GRADE approach to public health interventions: an empirical study**. *BMC public health* 2013, **13**(1):9.
62. Schünemann HJ: **Methodological idiosyncracies, frameworks and challenges of non-pharmaceutical and non-technical treatment interventions**. *Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen* 2013, **107**(3):214-220.
63. Pearce NE, Zahm SH, Andersen A, Antó i Boqué JM, Cardis E, Grimsrud TK, Kjaerheim K, Kogevinas M, Porta Serra M: **IARC monographs: 40 years of evaluating carcinogenic hazards to humans**. 2015.
64. Morgan RL, Thayer KA, Guyatt G, Blain R, Eftim S, Ross P, Santesso N, Holloway AC, Schunemann HJ: **Assessing the Usability of ACROBAT-NRSI for Studies of Exposure and Intervention in Environmental Health Research**. In: *Cochrane Colloquium.* Vienna, Austria; 2015.
65. Bilotta GS, Milner AM, Boyd IL: **Quality assessment tools for evidence from environmental science**. *Environmental Evidence* 2014, **3**(1):1-14.
66. Judson R, Richard A, Dix DJ, Houck K, Martin M, Kavlock R, Dellarco V, Henry T, Holderman T, Sayre P: **The toxicity data landscape for environmental chemicals**. *Environ Health Perspect* 2009, **117**(5):685-695.
67. EPA: **EPA Announces Actions to Address Chemicals of Concern, Including Phthalates**. 2009.
68. Tice RR, Austin CP, Kavlock RJ, Bucher JR: **Improving the human hazard characterization of chemicals: a Tox21 update**. *Environ Health Perspect* 2013, **121**(7):756-765.
69. Kavlock R, Chandler K, Houck K, Hunter S, Judson R, Kleinstreuer N, Knudsen T, Martin M, Padilla S, Reif D: **Update on EPA's ToxCast program: providing high throughput decision support tools for chemical risk management**. *Chemical research in toxicology* 2012, **25**(7):1287-1302.

70. IARC: **IARC: Diesel engine exhaust carcinogenic**. *Press release* 2012(213).
71. **HEALTH EFFECTS OF DIESEL EXHAUST: A fact sheet by Cal/EPA's Office of Environmental Health Hazard Assessment and the American Lung Association** [http://oehha.ca.gov/public_info/facts/dieselfacts.html]
72. **Tools & Resources Regulatory Standards** [http://www.epa.gov/cleandiesel/reg-prog.htm]
73. **Design for the Environment Alternatives Assessments** [http://www2.epa.gov/saferchoice/design-environment-alternatives-assessments]

Figures

Figure 1.

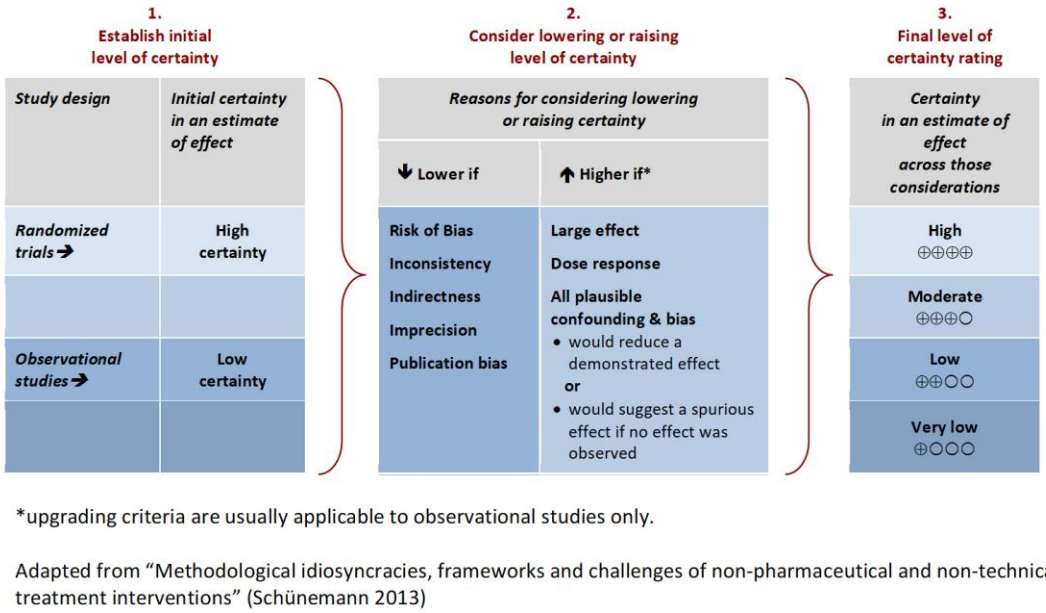


Figure 1. GRADE's approach to developing certainty ratings across a body of evidence for each outcome based on a systematic review and across outcomes (lowest quality across the outcomes critical for decision-making).

CHAPTER 3. EVALUATION OF THE RISK OF BIAS IN NON-RANDOMIZED STUDIES OF INTERVENTIONS (ROBINS-I) AND THE 'TARGET EXPERIMENT' CONCEPT IN STUDIES OF EXPOSURES: RATIONAL AND PRELIMINARY INSTRUMENT DEVELOPMENT

PREFACE TO CHAPTER 3

Chapter 3. *Evaluation of the Risk of Bias in Non-randomized Studies of Interventions (ROBINS-I) and the 'target experiment' concept in studies of exposures: rational and preliminary instrument development* was submitted to Environment International on 15 December 2017 and, as of 21 December 2017, is currently under review.

Evaluation of the Risk of Bias in Non-randomized Studies of Interventions (ROBINS-I) and the 'target experiment' concept in studies of exposures: rational and preliminary instrument development

Author list

Rebecca L. Morgan [a]; Kristina A. Thayer [b]; Nancy Santesso [a]; Alison C. Holloway [c]; Robyn Blain [d]; Sorina E. Eftim [d]; Alexandra E. Goldstone [d]; Pam Ross [d]; Holger J. Schünemann [a,e]

Affiliations

[a] Department of Health Research Methods, Evidence, and Impact (formerly the Department of Clinical Epidemiology & Biostatistics), McMaster University, Health Sciences Centre, Room 2C14, 1280 Main Street West, Hamilton, ON L8S 4K1 Canada

morganrl@mcmaster.ca, santesna@mcmaster.ca, schuneh@mcmaster.ca

[b] Integrated Risk Information System (IRIS) Division, National Center for Environmental Assessment (NCEA), Office of Research and Development, US Environmental Protection Agency, Building B (Room 211i), Research Triangle Park, NC USA 27711.

thayer.kris@epa.gov

[c] Department of Obstetrics and Gynecology, McMaster University, Health Sciences Centre, Room 3N52A, 1280 Main Street West, Hamilton, ON L8S 4K1 Canada.

hollow@mcmaster.ca

[d] ICF, 9300 Lee Highway, Fairfax, VA 22031 USA. Robyn.Blain@icf.com, Pam.Ross@icf.com, Ali.Goldstone@icf.com, Sorina.Eftim@icf.com

[e] Department of Medicine, McMaster University, Health Sciences Centre, Room 2C14, 1280 Main Street West, Hamilton, ON L8S 4K1 Canada. schuneh@mcmaster.ca

Corresponding author: Holger J. Schünemann. Department of Health Research Methods, Evidence, and Impact, Health Sciences Centre, Room 2C14, 1280 Main Street West, Hamilton, ON L8S 4K1 Canada. schuneh@mcmaster.ca.

Conflict of interest

The authors declare they have no competing financial interests with respect to this manuscript, or its content, or subject matter.

Abstract

Assessing the risk of bias (RoB) of individual studies is a critical part of the overall process used to determine the certainty of evidence from non-randomized studies (NRS) of potential health effects from environmental exposures. The recently released RoB in NRS of Interventions (ROBINS-I) instrument has undergone careful development and piloting on NRS of health interventions. Using the fundamental design of ROBINS-I, which includes evaluating RoB against an ideal target trial, we explored developing a version of the instrument to evaluate RoB in exposure studies. During three sequential rounds of assessment, two or three raters (evaluators) independently applied ROBINS-I to studies from two systematic reviews and one case-study protocol evaluating the relationship between environmental exposures and health outcomes. Feedback from raters, methodologists, and topic-specific experts in the field of environmental health research informed modifications to the instrument. We identified the following areas of distinction for the modified instrument: the process, formulating the target experiment, and evaluating exposure misclassification. The nature of these methodological changes facilitates RoB assessment of NRS of exposures in the environmental health field with structured evidence-synthesis frameworks, such as Grading of Recommendations Assessment, Development and Evaluation (GRADE).

Keywords (6): risk of bias; environmental health; GRADE; non-randomized studies;

environmental exposure; ROBINS

Highlights

- The adapted RoB instrument for NRS of exposures instrument reflects modifications suggested from an evaluation of a recently released instrument to assess RoB for health interventions (ROBINS-I).
- Authors can use the RoB instrument for NRS of exposures to evaluate the RoB of individual studies and across studies of environmental or occupational exposures by using the concept of the target experiment as a point of reference.
- As for systematic reviews of interventions, the RoB instrument for NRS of exposures used in conjunction with Grading of Recommendations Assessment, Development and Evaluation (GRADE) begins with a rating of 'High' but typically requires rating down for confounding and other biases unless lack of bias can be carefully justified.

Abbreviations

ACROBAT-NRSI: A Cochrane Risk of Bias Tool – for Non-randomized Studies of Interventions

BPA: bisphenol-A

CI: confidence interval

CoE: Certainty of evidence

EFSA: European Food Safety Authority

EPA: Environmental Protection Agency

GRADE: Grading of Recommendations Assessment, Development, and Evaluation

IRIS: Integrated Risk Information System

NOS: Newcastle-Ottawa Scale

NRS: Non-randomized studies

OHAT: Office of Health Assessment and Translation

ORoC: Office of the Report on Carcinogen

PBDE: polybrominated diphenyl ethers

PECO: population, exposure, comparator, outcome

PFOA: perfluorooctanoic acid

PICO: population, intervention, comparator, outcome

PME: particulate matter exposure

RCT: randomized controlled trial

RoB: Risk of bias

ROBINS-I: Risk of Bias in Non-randomized Studies of Interventions

T4: thyroid hormone thyroxine

TSH: thyroid simulation hormones

1. Introduction

Assessing the risk of bias (RoB) of individual studies is a critical part of the overall process used to determine the certainty of evidence from non-randomized studies (NRS) of potential health effects from environmental exposures. RoB, also called internal validity or limitations in the detailed design or execution, of the studies included in a review contributes to the overall evaluation of the certainty or quality of evidence.

The Risk of Bias in NRS of Interventions (ROBINS-I) instrument, formerly named A Cochrane Risk of Bias Tool for Non-randomized Studies of Interventions (ACROBAT-NRSI) was released in 2016 to assess RoB in non-randomized (i.e., observational) studies (NRS) of health interventions [1, 2]. This instrument allows users to identify and assess RoB in NRS that evaluate the effects of one or more interventions at the individual outcome level. ROBINS-I, based on the Cochrane RoB instrument for randomized controlled trials (RCTs), facilitates a structured comparison of NRS to the gold standard represented by the RCT or target trial [3]. Signaling questions in the ROBINS-I instrument prompt raters to assess RoB in domains of: 1) bias due to confounding, 2) bias in selection of participants into the study, 3) bias in classification of interventions, 4) bias due to departures from intended interventions, 5) bias due to missing data, 6) bias in measurement of outcomes, and 7) bias in selection of reported results. Additional elements of the ROBINS-I instrument include an optional component to judge the direction of the bias for each domain. ROBINS-I was designed to inform the rating of the certainty of a body of evidence (CoE) regarding health interventions within the Grading

of Recommendations Assessment, Development, and Evaluation (GRADE) framework [2].

Use of the target trial concept in ROBINS-I has impact on the application of the GRADE framework for assessing the CoE that include NRS [4]. The GRADE framework allows reviewers to systematically and transparently assess the CoE in a body of evidence to inform decision making [5]. GRADE considers the following factors as decreasing CoE: RoB, inconsistency, indirectness, imprecision, and publication bias. In addition, three factors increase CoE: large magnitude of effect, dose-response gradient, and opposing residual confounding. Within GRADE, because of the balance of prognostic factors that RCTs provide, RCTs start at 'High' initial CoE and NRS start at 'Low' initial CoE; however, the introduction of instruments using a standardized comparison for RoB assessment, such as ROBINS-I, eliminate the requirement for labelling a body of evidence as high or low certainty based on study design [4].

Since the ROBINS-I instrument was developed for health intervention studies, which often involve intentional 'exposures', the usefulness of ROBINS-I to evaluate RoB in studies of exposures to environmental chemicals, which are typically unintended exposures, is unclear. Conceptually, there was reason to expect that ROBINS-I would extrapolate to environmental exposure because the RoB domains overlap with instruments used in the field, such as the Newcastle-Ottawa Scale (NOS) and those developed by the Environmental Protection Agency (EPA), European Food Safety Authority (EFSA), the National Toxicology Program's Office of Health Assessment and Translation (OHAT) and Office of the Report on Carcinogen (ORoC), and University of

California in San Francisco's (UCSF) Navigation Guide [6-12]. Although the domain content is broadly similar across these instruments, they focus on study-design features, include items related to the sensitivity of the study, and not all use signaling questions [13, 14]. Most importantly, the ideal study design concept is not explicitly outlined. As GRADE considers the impact of the use of ROBINS-I on the conceptualization of the CoE, development of a version that allows assessment of studies dealing with exposures will facilitate harmonization of rating NRS of interventions and environmental or occupational exposures in the context of GRADE.

In this paper, we present a RoB instrument for NRS of exposures, developed from a series of pilot tests and external feedback to ROBINS-I, to evaluate reviews of environmental exposure studies. We highlight the common methodological challenges and considerations experienced when we applied the ROBINS-I instrument to studies of unintentional exposures.

2. Development of a RoB instrument to evaluate studies of exposure

2.1. Methods

We evaluated the ROBINS-I instrument in environmental studies of exposure by applying it to two existing systematic reviews and one case study protocol during three sequential rounds of assessment [15-17] (Appendices A & B). We identified facilitators and barriers to implementation of the instrument. Feedback from a group of raters (evaluators) conducting the pilot testing, as well as methodologists and topic-specific

experts in the field of environmental health research, informed modifications to the ROBINS-I instrument. During pilot testing three raters independently evaluated each study and discussed judgments to reach a consensus rating. To inform modifications to this RoB instrument for NRS of exposures, the raters were familiar with epidemiological methods; however, to model real-world RoB assessment we also selected raters who did not have content-specific knowledge of the exposure or outcomes. Raters had access to topic-specific experts for guidance and to improve understanding throughout the process.

ROBINS-I served as the platform for our initial assessment; however, it became clear that modifications were required to conduct RoB in NRS of exposure, so we subsequently referred to the modified instrument as a RoB instrument for NRS of exposures. A steering group of key investigators (RM, KT, AH, NS, HS) made decisions regarding whether or not to modify the instrument based on user experience during pilot testing.

Three rounds of testing and feedback from methodologists and topic-specific experts suggested semantic and conceptual modifications to facilitate understanding and enhance usefulness of the instrument in environmental health (Appendix C). Steering group members agreed both semantic and conceptual modification to the ROBINS-I instrument were sufficient to necessitate a distinct instrument for assessing RoB in NRS of exposure (Appendix D). We posted this preliminary version of a RoB instrument for NRS of exposures on the University of Bristol website in 2017, so that interested organizations could pilot and provide feedback for further development [18].

3. Methodological distinctions between assessing RoB in NRS of interventions and in NRS of exposures

3.1. Conducting the RoB assessment

A RoB instrument for NRS of exposures has three stages: 1) clarify the review question and identify confounders and co-exposures; 2) describe a target trial/experimental version of the study; and 3) evaluate the study. In Stage I, the review group describes the question of the review: the population, exposure, comparators, outcome (PECO), as well as general information regarding confounders, co-exposures or interventions, and assessment of the exposure and health outcome. All study outcomes of interest are included in the 'O' field. In Stage II, for each eligible study, the review group describes the target experiment, including confounders and co-interventions that could have occurred in the study. This includes identifying the specific outcome of interest that raters will evaluate with the instrument, as both instruments are designed to evaluate the RoB of a study for a specific outcome. Lastly, in Stage III, raters evaluate each study using the RoB signaling questions and make domain-level judgments.

At each stage in the process, the involvement of topic-specific experts is paramount to identify confounders and co-interventions, nuances of the exposure and outcome measurements, as well as review the eligibility of studies, and the accuracy and completeness of rationales provided in response to the signaling questions.

3.2. The target experiment

ROBINS-I presents a distinguishing feature from previously published RoB instruments for evaluating NRS: the target trial [1]. Each review question is formulated to emulate a hypothetical pragmatic RCT. Evaluations of the target trial within ROBINS-I by topic-specific experts and methodologists revealed initial confusion and criticism when applied to studies of environmental exposure, namely that the trial may not only emulate an RCT, but may also emulate an animal experiment. For pilot testing, we renamed the 'intervention' of the population, intervention, comparator, outcome (PICO) question in ROBINS-I to 'experimental exposure' to maintain consistency with the transition to the PECO. Thus, with input from ROBINS-I developers, we renamed 'target trial' to 'target experiment' to emphasize that the unintentional exposures assessed by the modified instrument could also be compared to a hypothetical animal experiment.

We also identified issues related to the exposure and comparator of the target experiment. The ROBINS-I instrument is used to evaluate NRS that compare outcomes in at least two groups (i.e. comparative studies). Primary studies should include an 'intervention group' and an alternate control or comparison, whether provided in the study or by the initial research question. The comparator could address thresholds, levels, durations, ranges, means, medians, or ranges of exposure, including an incremental increase in exposure [19]. When using the ROBINS-I instrument, in Stage I the assessor describes the PECO question.

3.3. Misclassification of the exposure

Environmental exposures can often be conceptualized as unintentional interventions, which poses a challenge for accurate ascertainment of the exposure. One fundamental

challenge is evaluating confidence in exposure characterization [14]. For example, how certain are we that an individual in the lower exposure group is correctly classified to that group? In addition, information about the timing of exposure is often unavailable. Within the RoB instrument for NRS of exposures, the rater collects two different types of information to characterize the exposure. Prior to answering the RoB signaling questions, in Stage I, the rater collects information about how applicable the measurement of the exposure is to the 'E' of the PECO question. When responding to questions about bias in each study, in Stage III, the rater collects information on the validity (i.e., the most robust exposure assessment methods) of the measures used and whether or not they distinguish between the exposed and comparative groups.

Raters who pilot tested ROBINS-I and early drafts of the RoB instrument for NRS of exposures conflated issues of applicability (i.e., generalizability) and RoB. By default, raters captured these concepts within their responses to the RoB signaling questions. In the RoB instrument for NRS of exposures, we maintain a distinction between whether the exposure was measured incorrectly among all persons in the study and whether the exposure was measured incorrectly among some persons placing them in the inappropriate exposure or comparator group. Stages I and II in the RoB instrument for NRS of exposures contain fields to capture information on the accuracy of the exposure measurement that can inform how directly the 'E' and 'C' in the PECO are measured. The domain dedicated to 'Misclassification of the Exposure' includes signaling questions regarding how the exposure status is defined (i.e. were participants allocated to the correct exposure categories) and the robustness of the exposure assessment methods

(i.e. were the methods used to measure the exposure conducted correctly). This distinction promotes integration into decision-making frameworks, such as GRADE, which assess RoB as one factor when considering the CoE [20].

We identified similar confusion when raters evaluated studies with a cross-sectional design. Cross-sectional design studies are common in the environmental health literature; at a single time point they allow measuring environmental exposures, health outcomes, as well as certain confounders, and may require less time to complete and be more feasible to conduct than other study designs [21]. However, cross-sectional designs are at particularly high risk of bias. Specific concerns of bias in cross-sectional studies include concerns about the ability to establish temporality of the potential association between exposure and outcome and the desirability of having multiple measurements of an exposure to assess stability of the exposure levels over time. Some exposures require measurement at multiple time points (i.e., non-persistent chemicals) to determine their presence and quantify levels in the body. The RoB instrument for NRS of exposures delineates such concerns by distinguishing the measurements needed to determine exposure accuracy in Stage I and the degree of RoB if such measures are applied incorrectly.

An additional concept not yet addressed within the development of the target experiment or measurement of the exposure relates to the uncertainty about a study's ability to detect a true effect, called study sensitivity [22]. This could happen when the levels of exposure do not include sufficient variation among subjects to allow detecting an effect. For PECO questions that explore if an association between different exposure

levels and outcomes exist, studies at lower RoB will include sufficient variation in exposure levels among subjects to detect potential associations.

4. Discussion

We developed an instrument to evaluate the RoB of environmental exposure studies as an ideal target experiment, building on the strong methodological foundation established by the ROBINS-I work group [1]. Recognizing the substantial differences when assessing the CoE related to the effects of exposures on outcomes compared with intentional interventions, systematic review authors and guideline developers can use the instrument to assess risk of bias of individual studies and across studies. Development of the final RoB instrument for NRS of exposures will facilitate the use of NRS of environmental, nutritional or occupational exposures in GRADE.

4.1. Strengths and weaknesses of the study

Strengths of this work include our undertaking of a multi-step process to determine the extent to which the ROBINS-I instrument is applicable to assessing the health impact of environmental exposures (Appendix B). To assess and improve the presentation and relevance (i.e. face validity) of the instrument, at least one topic-specific expert in the field of interest populated Stage I during an iterative process. Topic-specific experts provided suggestions for improved applicability of the instrument in three ways during the evaluation and instrument adaptation process: 1) by providing background information for raters when applying ROBINS-I and the modified instrument; 2) by providing guidance on additional questions to add to ROBINS-I specific to environmental

exposure; and 3) by performing additional piloting of the modified instrument. We found that information specific to identifying and measuring the exposures and outcomes was important when responding to the corresponding signaling questions.

While the same raters applied the instrument to individual studies within the systematic reviews, we did not formally test or calculate reliability of responses. This would require a final version of the modified instrument. However, based on the narrative responses to the signaling questions, even with the initial adjustments made to the ROBINS-I instrument to address exposures, raters reported misunderstanding the concepts in the questions and the information in the studies. Modifications to the instrument and instructions from the three rounds of pilot testing and external feedback improved understanding.

4.2. Strengths and weaknesses in relation to other instruments

Through pilot-testing feedback and external consultation, we evaluated the face validity of the RoB instrument for NRS of exposures. In addition, the modified instrument, adapted from ROBINS-I, has both similarities and differences when compared with other RoB instruments used to assess studies of environmental exposure [14]. Domains that assess aspects of bias due to confounding, selection of participants, measurement of exposure, intended exposure, missing data, measurement of outcomes, and reported results are also reflected in other instruments (NOS, EPA, EFSA, Navigation Guide, OHAT, and ORoC instruments). However, the modified instrument remains distinct in its assessment of how well individual studies emulate the target experiment, which impacts evaluation of each bias domain and overall RoB judgments [1, 14]. Additionally,

the modified instrument maintains a distinction between the evaluation of RoB and of other factors pertinent to assessing the utility of a study, in particular the concept of study sensitivity [22]. Additional work remains on how to best conceptualize study sensitivity both within the RoB instrument for NRS of exposures, as well as within the GRADE CoE framework.

While RoB instruments evaluating RCTs or intentional interventions assume that researchers have control over or knowledge of the initiation of the intervention, this is not frequently the case for studies of exposures. RoB instruments evaluating exposure need to confirm the validity of the exposure measurement, the applicability of the measurement used in each study, and the ability of the exposure measurement to distinguish a difference in the outcome.

4.3. Implications for researchers and policymakers

When using evidence to inform decision making, the RoB instrument for NRS of exposures allows systematic-review authors and guideline developers to evaluate NRS with the target experiment as a reference point. Use of the ideal study design concept may be implicit in the instructions for some of the instruments, but it is not as explicitly outlined. Developers preferring to focus on assessing RoB based on domains rather than strongly based on study design labels, which allows for a more nuanced assessment, may favor the RoB instrument for NRS of exposures. Using the RoB instrument for NRS of exposures also facilitates assessment of the overall CoE because of its integration with GRADE. When the RoB instrument for NRS of exposures is used with GRADE, all studies start at a high certainty rating and the detailed assessment of the GRADE

domains, including risk of bias, then determines the final overall certainty. Planning for the application of the modified instrument should recognize the resources needed to conduct the RoB assessment as intended, such as time demands and topic-specific expertise.

4.4. Unanswered questions and future research

There is still a need to further assess the reliability and validity of the instrument; however, similar to other RoB instruments (e.g., ROBINS-I and the Cochrane RoB instrument for RCTs), much of that information will be gained by external application and feedback. Future studies could assess the reliability of raters' responses to the modified instrument in comparison to other RoB instruments used in environmental health. Studies also need to evaluate the validity of the modified instrument to assess RoB in a variety of exposure scenarios, including occupational exposures and exposures characterized by techniques other than biomonitoring. Similar to methods used for the development of this instrument, the involvement of topic-specific experts and iterative rounds of pilot testing will be needed. Comprehensive guidance with examples is needed for raters and for decision makers using the output from the RoB instrument for NRS of exposures.

The RoB instrument for NRS of exposures requires an independent control or comparative group to provide an evaluation that emulates a target experiment. Guidance for the rationale and approach to PECO formulation is forthcoming [23]. Piloting of ROBINS-I and the modified instrument identified continued confusion on the topic of RoB and other factors related to assessing the CoE [13]. Some consider one's

ability to determine an exposure's true effect (e.g., 'study sensitivity'), as distinct from RoB and issues of indirectness but this depends on the PECO question that is asked in the systematic review [22]. In this instrument, fields to address study sensitivity could be added to Stages I or II in the instrument, where the concept of indirectness is addressed.

5. Conclusions

We evaluated the application of the ROBINS-I instrument to NRS of exposures by applying it to two existing systematic reviews and one case-study protocol. Based on a three-stage, pilot-testing study that involved numerous raters, topic-specific experts, and collaboration with the original instrument developers, we modified an existing RoB instrument for evaluation of environmental studies of exposure. Modifications made to the ROBINS-I instrument to tailor it to studies of environmental exposure increased understanding and application of the instrument. The modifications made to the instrument were important enough to recommend an instrument distinct from ROBINS-I for NRS of exposure. This RoB instrument for NRS of exposures can serve as a standardized, transparent, and rigorous instrument for evaluating RoB of environmental exposure studies. It lends itself to the use in the context of GRADE to assess the overall certainty in a body of evidence, but users should be aware of the special consideration around the initial CoE.

6. Acknowledgements

We would like to acknowledge contributions from Juleen Lam, Danielle Mandrioli, Kavita Singh for their review and application of the modified ROBINS-I instrument; and Ruth Lunn and Gloria Cooper, for their input when comparing signaling questions with other instruments used for RoB assessment in environmental health; and Julian Higgins and Jonathan Sterne, for comments on the final instruments.

7. Funding Sources

This research was supported by the Intramural Research Program of the National Institute of Environmental Health Sciences and the GRADE Centre at the McMaster University.

8. References

1. Sterne JA, Hernan MA, Reeves BC, Savovic J, Berkman ND, Viswanathan M, Henry D, Altman DG, Ansari MT, Boutron I *et al*: **ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions**. *BMJ* 2016, **355**:i4919.
2. **ACROBAT-NRSI: A Cochrane Risk Of Bias Assessment Tool for Non-Randomized Studies of Interventions. Accessed 24 September 2014.** [https://sites.google.com/site/riskofbiastool/]
3. Cochran WG, Chambers SP: **The planning of observational studies of human populations**. *Journal of the Royal Statistical Society Series A (General)* 1965, **128**(2):234-266.
4. Schünemann H, Cuello C, Akl EA, Mustafa R, Meerpohl J, Thayer K, Morgan RL, Gartlehner G, Kunz R, Katikireddi S *et al*: **GRADE Guidelines: 18. How tools to assess risk of bias in non-randomized studies should be used to rate the certainty of a body of evidence** *Journal of clinical epidemiology* Under review.
5. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, Schunemann HJ, Group GW: **GRADE: an emerging consensus on rating quality of evidence and strength of recommendations**. *BMJ* 2008, **336**(7650):924-926.
6. LaKind JS, Sobus JR, Goodman M, Barr DB, Furst P, Albertini RJ, Arbuckle TE, Schoeters G, Tan YM, Teeguarden J *et al*: **A proposal for assessing study quality: Biomonitoring, Environmental Epidemiology, and Short-lived Chemicals (BEES-C) instrument**. *Environ Int* 2014, **73**:195-207.
7. Koustas E, Lam J, Sutton P, Johnson PI, Atchley DS, Sen S, Robinson KA, Axelrad DA, Woodruff TJ: **The Navigation Guide - evidence-based medicine meets environmental health: systematic review of nonhuman evidence for PFOA effects on fetal growth**. *Environ Health Perspect* 2014, **122**(10):1015-1027.
8. NRC (National Research Council): **Review of EPA's Integrated Risk Information System (IRIS) Process (http://www.nap.edu/catalog.php?record_id=18764) [accessed 1 January 2015]**. 2014.
9. NTP (National Toxicology Program): **Handbook for Conducting a Literature-Based Health Assessment Using Office of Health Assessment and Translation (OHAT) Approach for Systematic Review and Evidence Integration. January 9, 2015 release. Available at http://ntp.niehs.nih.gov/go/38673.** 2015.
10. NTP (National Toxicology Program): **Handbook for Preparing Report on Carcinogens Monographs - July 2015. Available at http://ntp.niehs.nih.gov/go/rochandbook**. 2015(January 3, 2017).
11. Wells G, Shea B, O'Connell D, Peterson J, Welch V, Losos M, Tugwell P: **The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses**. In.; 2000.
12. Authority EFS: **Tools for critically appraising different study designs, systematic review and literature searches**. *EFSA Supporting Publication* 2015, **12**(7).
13. **Methods Commentary: Risk of Bias in Cohort Studies.** [https://distillercer.com/resources/methodological-resources/risk-of-bias-commentary/]

14. Rooney AA, Cooper GS, Jahnke GD, Lam J, Morgan RL, Boyles AL, Ratcliffe JM, Kraft AD, Schünemann HJ, Schwingl P: **How credible are the study results? Evaluating and applying internal validity tools to literature-based assessments of environmental health hazards**. *Environment international* 2016.
15. Johnson PI, Sutton P, Atchley DS, Koustas E, Lam J, Sen S, Robinson KA, Axelrad DA, Woodruff TJ: **The Navigation Guide - evidence-based medicine meets environmental health: systematic review of human evidence for PFOA effects on fetal growth**. *Environ Health Perspect* 2014, **122**(10):1028-1039.
16. Thayer K, Rooney A, Boyles A, Holmgren S, Walker V, Kissling G, U.S. Department of Health and Human Services: **Draft protocol for systematic review to evaluate the evidence for an association between bisphenol A (BPA) exposure and obesity**. *National Toxicology Program* 2013.
17. Zhao X, Wang H, Li J, Shan Z, Teng W, Teng X: **The Correlation between Polybrominated Diphenyl Ethers (PBDEs) and Thyroid Hormones in the General Population: A Meta-Analysis**. *PLoS One* 2015, **10**(5):e0126989.
18. **The ROBINS-E tool (Risk Of Bias In Non-randomized Studies - of Exposures)** [https://www.bristol.ac.uk/population-health-sciences/centres/cresyda/barr/riskofbias/robins-e/]
19. Morgan RL, Thayer K, Whaley P, Schünemann H: **Identifying the PECO: A framework for formulating good questions to explore the association of environmental and other exposures with health outcomes**. Unpublished.
20. Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, Montori V, Akl EA, Djulbegovic B, Falck-Ytter Y *et al*: **GRADE guidelines: 4. Rating the quality of evidence--study limitations (risk of bias)**. *J Clin Epidemiol* 2011, **64**(4):407-415.
21. Levy BS: **Occupational and environmental health: recognizing and preventing disease and injury**: Lippincott Williams & Wilkins; 2006.
22. Cooper GS, Lunn RM, Agerstrand M, Glenn BS, Kraft AD, Luke AM, Ratcliffe JM: **Study sensitivity: Evaluating the ability to detect effects in systematic reviews of chemical exposures**. *Environ Int* 2016, **92-93**:605-610.
23. Morgan RL, Thayer KA, Whaley P, Schünemann HJ: **Identifying the PECO: Strategies for formulating decision-making questions for environmental health**. Unpublished.

Appendices

Appendix A. Characteristics of systematic reviews assessed using the ROBINS-I instrument.

Pilot Round	Title (Reference)	Exposure	Outcome	Number of Studies	Number of raters per study	Study design (n)
Round 1	Draft protocol for systematic review to evaluate the evidence for an association between bisphenol A (BPA) exposure and obesity[1]	BPA	Overweight and obesity	14	2	Cohort (2) Cross-sectional (12)
Round 2	The Navigation Guide-Evidence-Based Medicine Meets Environmental Health: Systematic Review of Human Evidence for PFOA Effects on Fetal Growth[2]	PFOA	Fetal growth (i.e., birth weight)	17	2	Cohort (7) Cross-sectional (10)
Round 3	The Correlation between Polybrominated Diphenyl Ethers (PBDEs) and Thyroid Hormones in the	PBDEs	Thyroid function as measured by thyroid simulation hormones (TSHs) or thyroid	17	3	Cohort (3) Case-control (1) Cross-sectional (13)

[1] Thayer K, Rooney A, Boyles A, Holmgren S, Walker V, Kissling G, U.S. Department of Health and Human Services: Draft protocol for systematic review to evaluate the evidence for an association between bisphenol A (BPA) exposure and obesity. National Toxicology Program 2013.

[2] Johnson PI, Sutton P, Atchley DS, Koustas E, Lam J, Sen S, Robinson KA, Axelrad DA, Woodruff TJ: The Navigation Guide-Evidence-Based Medicine Meets Environmental Health: Systematic Review of Human Evidence for PFOA Effects on Fetal Growth. Environmental health perspectives 2014.

	General Population: A Meta-Analysis[3]		hormone thyroxine (T4)			

BPA: bisphenol A; PBDE: polybrominated diphenyl ethers; PFOA: perfluorooctanoic acid.

[3] Zhao XM, Wang HL, Li J, Shan ZY, Teng WP, Teng XC: The Correlation between Polybrominated Diphenyl Ethers (PBDEs) and Thyroid Hormones in the General Population: A Meta-Analysis. Plos One 2015, 10(5).

Appendix B. Detailed methods of the evaluation of ROBINS-I and development of the RoB instrument for NRS of exposures.

Methods

Approach

We evaluated the ROBINS-I instrument in occupational and environmental studies of unintentional exposure by applying it to two existing systematic reviews and one draft case study protocol. We focused on identifying benefits and barriers to implementation of the instrument. Feedback from a group of raters (evaluators) conducting the pilot testing, as well as methodologists and topic-specific experts in the field of environmental health research, informed modifications to the ROBINS-I instrument. ROBINS-I served as the platform for our initial assessment; however, when it became clear that certain modifications were required to conduct RoB in NRS of exposure, we referred to it as the modified instrument. A steering group of key investigators (RM, KT, AH, NS, HS) made decisions regarding whether or not to modify the instrument based on these findings.

Instruments

Initially released as ACROBAT-NRSI in 2014 and renamed as ROBINS-I in 2016, this study used both iterations of the instrument when assessing understanding and applicability to environmental exposure studies[4]. We will refer to either ACROBAT-NRSI or ROBINS-I

[4] Sterne JA, Hernan MA, Reeves BC, Savovic J, Berkman ND, Viswanathan M, Henry D, Altman DG, Ansari MT, Boutron I et al: ROBINS-I: a tool for assessing risk of bias in non-

based on the version of the instrument that was used during that stage in the assessment.

The ACROBAT-NRSI/ROBINS-I instrument contains 30 signaling questions across seven bias domains to assist reviewers in determining whether an individual study is at low, moderate, serious, or critical RoB (www.riskofbias.info)[5]. ROBINS-I focuses on studies of interventions in cohort and case-control study designs, with plans to explore modifications in future for other study types such as cross-sectional studies. The seven bias domains are: 1) Bias due to confounding, 2) Bias in selection of participants into the study, 3) Bias in classification of interventions, 4) Bias due to departures from intended exposures, 5) Bias due to missing data, 6) Bias in measurement of outcomes, and 7) Bias in selection of the reported result. Using this instrument, raters can determine domain-level judgments within a study and study-level judgments about RoB based on the seven domain-level judgments. Signaling question response options include 'Yes', 'Probably yes', 'Probably no', 'No', and 'No information', and are complemented by free text fields to capture response judgments. Raters use the signaling question and free-text responses to make domain-level judgments about RoB. Domain- and study-level response options include 'Low', 'Moderate', 'Serious', and 'Critical' RoB. The individual

randomised studies of interventions. BMJ 2016, 355:i4919. ACROBAT-NRSI: A Cochrane Risk Of Bias Assessment Tool for Non-Randomized Studies of Interventions. Accessed 24 September 2014. [https://sites.google.com/site/riskofbiastool/].

[5] Sterne JA, Hernan MA, Reeves BC, Savovic J, Berkman ND, Viswanathan M, Henry D, Altman DG, Ansari MT, Boutron I et al: ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016, 355:i4919. ACROBAT-NRSI: A Cochrane Risk Of Bias Assessment Tool for Non-Randomized Studies of Interventions. Accessed 24 September 2014. [https://sites.google.com/site/riskofbiastool/].

study-level RoB is typically taken from the most severe of the domain-level judgments, unless the rater feels that the individual study should be rated as having greater RoB than that based on several affected domains. Domain-level responses across a body of evidence (across studies) allow an assessment of how much the domain-level RoB judgments may contribute to the trustworthiness of the entirety of evidence. It is recommended that studies with an overall RoB of 'Critical risk' not be included in a meta-analysis; however, that decision should be made considering the totality of the evidence[6].

Preparation for an evaluation using this instrument includes populating both a project- and an individual study-level protocol[6]. For each research question, raters complete one project-level protocol, identifying their target randomized trial research question. The target randomized trial research question identifies the population, intervention, comparison, and outcomes of interest. Based on this target trial, raters identify the nature of the target comparison (i.e., effect of interest), potential confounders and the relationship between them and the confounding domains for the research project. It also includes addressing possible co-interventions that could have an impact on the study outcomes, and the result(s) being assessed. For each individual study eligible to answer the review question, reviewers complete a study-level protocol. Text fields in the study-level protocol reflect those in the project-level protocol, to facilitate the abstraction of information from each individual study to determine generalizability and

[6] Sterne JA, Hernan MA, Reeves BC, Savovic J, Berkman ND, Viswanathan M, Henry D, Altman DG, Ansari MT, Boutron I et al: ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016, 355:i4919.

applicability to answering the project-level research question. Raters extract information to assess whether or not confounders and co-interventions identified as critical were addressed in the individual study and whether the individual study identified additional confounders or co-interventions.

Selection of raters and topic-specific experts

Raters

We selected raters (RB, SE, AG, and PR) with master's and doctoral degrees, training in epidemiological methods, and at least four years (range 4-13 years) of experience as evaluators of epidemiological studies. While the raters did not necessarily have topic-specific expertise on the environmental exposures in the selected systematic reviews, they had access to topic-specific experts and other resources for consultation throughout the project. Raters initially received training materials, which included Stage I, Stage II, and the abstraction instrument per the ACROBAT-NRSI handbook, and supplemental information when the ROBINS-I iteration was released[7].

Case topic-specific experts

We selected topic-specific experts (JL, KS, and KT), familiar with the exposures and outcomes assessed during the three rounds of piloting, to provide the required

[7] Sterne JA, Hernan MA, Reeves BC, Savovic J, Berkman ND, Viswanathan M, Henry D, Altman DG, Ansari MT, Boutron I et al: ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016, 355:i4919. ACROBAT-NRSI: A Cochrane Risk Of Bias Assessment Tool for Non-Randomized Studies of Interventions. Accessed 24 September 2014. [https://sites.google.com/site/riskofbiastool/].

background information for completion of the evaluation using the ACROBAT-NRSI instrument, including about confounders and possible co-exposures. These PhD-level topic-specific experts had published articles on and had first-hand knowledge of the exposure and features important to the exposures and health outcomes of interest (i.e., BPA and obesity, PFOA and birth weight, and PBDE and thyroid function).

Systematic reviews selected for pilot testing

We assessed the utility of ACROBAT-NRSI by piloting the instrument on all primary studies included in two previously published systematic reviews and studies identified from a draft case study protocol developed by OHAT as part of its early efforts to implement systematic review[8]. We selected previously published systematic reviews and a draft case study protocol that presented both persistent and non-persistent chemicals, as well as included primary studies featuring a variety of NRS designs (cohort, case-control, and cross-sectional). Using ACROBAT-NRSI, raters evaluated studies identified in two systematic reviews and one draft case study protocol of environmental epidemiological studies: 1) exposure to bisphenol A (BPA) and its association with obesity; 2) developmental exposure to perfluorooctanoic acid (PFOA) and its effect on

[8] Thayer K, Rooney A, Boyles A, Holmgren S, Walker V, Kissling G, U.S. Department of Health and Human Services. Draft protocol for systematic review to evaluate the evidence for an association between bisphenol A (BPA) exposure and obesity. National Toxicology Program 2013. Johnson PI, Sutton P, Atchley DS, Koustas E, Lam J, Sen S, Robinson KA, Axelrad DA, Woodruff TJ: The Navigation Guide-Evidence-Based Medicine Meets Environmental Health: Systematic Review of Human Evidence for PFOA Effects on Fetal Growth. Environmental health perspectives 2014. Zhao XM, Wang HL, Li J, Shan ZY, Teng WP, Teng XC: The Correlation between Polybrominated Diphenyl Ethers (PBDEs) and Thyroid Hormones in the General Population: A Meta-Analysis. Plos One 2015, 10(5).

fetal growth; and 3) exposure to polybrominated diphenyl ethers (PBDEs) and its effect on thyroid function (Appendix A). Each of the reviews represented a collection of 14-19 studies, most of which were cross-sectional in design but also included several cohort studies.

Evaluation of selected systematic reviews

For the first and second rounds of user testing, which informed initial revisions to the instrument, two raters independently responded to the signaling questions and provided domain- and study-level RoB judgments according to the ACROBAT-NRSI instrument from each study within the selected systematic reviews on BPA and PFOA into Microsoft Excel. A third rater reviewed the results, established consensus, and determined overall RoB for each study. In the third round of user testing, the three raters independently applied the modified instrument to a systematic review looking at the impact of PBDEs on thyroid function. The three raters then agreed on overall RoB for each study.

Our rating protocol, developed for each review, identified the hypothetical (i.e., target) randomized trial, potential confounders, and possible co-exposures of interest. Initially, raters identified pre-specified chemical confounders and possible co-exposures related to the health outcomes. We used sources such as the PhenX Toolkit (https://www.phenxtoolkit.org/) to identify key confounders for the health outcomes[9].

[9] Hamilton CM, Strader LC, Pratt JG, Maiese D, Hendershot T, Kwok RK, Hammond JA, Huggins W, Jackman D, Pan H: The PhenX Toolkit: get the most from your measures. American journal of epidemiology 2011, 174(3):253-260.

Topic-specific experts provided guidance to address raters' unfamiliarity with the topic of each systematic review. When raters recognized additional confounders or co-exposures mentioned in the studies, these were added to the protocol; all studies were then re-evaluated so that raters considered the most comprehensive lists of confounders and co-exposures.

In the three rounds of pilot testing, raters received a form to identify and document barriers and facilitators to the use of the ROBINS-I in studies of environmental health. Also, we asked raters to provide descriptions of their understanding of each signaling question in the ROBINS-I instrument to identify areas requiring additional clarity and/or rewording. When deciding to modify ROBINS-I for the subsequent rounds of pilot testing, we considered modifications suggested by raters: for example, repeated misunderstanding of specific signaling questions over the multiple rounds of pilot testing.

Instrument evaluation and refinement

In addition to the three rounds of pilot testing, the steering committee consulted with other topic-specific experts in environmental health, as well as developers from the ACROBAT-NRSI/ROBINS-I instruments using a modified Delphi (group decision-making) technique conducted in each of the three rounds[10]. First, topic-specific experts from government and non-governmental organizations (GC, JL, RL) weighed in on the initial

[10] Hsu C-C, Sandford BA: The Delphi technique: making sense of consensus. Practical assessment, research & evaluation 2007, 12(10):1-8.

ACROBAT-NRSI items and identified similarities and differences with current instruments applied to studies of environmental exposures. To identify salient items, topic-specific experts compared RoB instruments used for evaluation of environmental studies used by the EPA, Navigation Guide, OHAT, and ORoC to determine common items that would identify typical classification errors in environmental risk[11].

Second, three case topic-specific experts (JL, DM, KS) in the field of environmental health all assessed one study from the review of BPA and obesity and one study from the review of PBDE and thyroid function using the modified ROBINS instrument and provided feedback on the signaling questions[12]. In addition, case topic-specific experts addressed responses to the signaling questions from raters to weigh in on the accuracy and comprehensiveness of their responses.

Lastly, chairs of the ACROBAT-NRSI/ROBINS-I instrument development work group (JH & JS) provided input on items added to tailor modifications made to evaluate studies of

[11] NRC (National Research Council): Review of EPA's Integrated Risk Information System (IRIS) Process (http://www.nap.edu/catalog.php?record_id=18764) [accessed 1 January 2015]. 2014. NTP (National Toxicology Program): Handbook for Preparing Report on Carcinogens Monographs - July 2015. Available at http://ntp.niehs.nih.gov/go/rochandbook. 2015(January 3, 2017). Johnson PI, Sutton P, Atchley DS, Koustas E, Lam J, Sen S, Robinson KA, Axelrad DA, Woodruff TJ: The Navigation Guide-Evidence-Based Medicine Meets Environmental Health: Systematic Review of Human Evidence for PFOA Effects on Fetal Growth. Environmental health perspectives 2014. Rooney AA, Boyles AL, Wolfe MS, Bucher JR, Thayer KA: Systematic review and evidence integration for literature-based environmental health science assessments. Environ Health Perspect 2014, 122(7):711-718.

[12] Carwile JL, Michels KB: Urinary bisphenol A and obesity: NHANES 2003–2006. Environmental research 2011, 111(6):825-830. Chevrier J, Harley KG, Bradman A, Gharbi M, Sjödin A, Eskenazi B: Polybrominated Diphenyl Ether (PBDE) Flame Retardants and Thyroid Hormone during Pregnancy. Environmental Health Perspectives 2010, 118(10):1444.

environmental exposure. As mentioned previously, our study incorporated updated versions of ROBINS-I into this pilot work on our modified instrument.

Data analysis

When discrepancies were identified during the first and second round of testing, the third reviewer discussed with the two raters to determine the basis for the discrepancy, i.e., confusion on the item or differences of opinion on the raters' observations. We discussed differences related to the clarity of the item and either reworded the item or provided additional guidance for the question, as necessary. Similarly, in the third round of user testing, all three raters that provided the review of the studies discussed and arrived at a consensus on the response to each instrument item and overall RoB.

Appendix C. Modifications made as a result of three rounds of pilot testing and external consultation.

Methods used during pilot-testing of ROBINS-I and subsequent modifications of the instrument	1) Development of Stage I: a. *A priori*, topic-specific experts of the environmental exposures of interest provided input to stage 1 of the instrument, identifying critical confounders, potential co-exposures, and identifying characteristics of the exposure and health outcome measurement accuracy, such as its persistence. b. Raters consulted a database on chemical and environmental exposures, the PhenX Toolkit (https://www.phenxtoolkit.org/), to identify potential confounders [19]. c. Topic-specific experts provided background information for raters when applying ROBINS-I or the modified instrument. 2) Completion of Stage II & III a. To improve reliability of responses, at minimum, two raters independently applied the instrument to each study in the systematic reviews, and compared and discussed their evaluations to reach consensus. b. Topic-specific experts performed additional piloting of the modified instrument.
Round 1: BPA and obesity	3) Replacement of the word 'intervention' with 'exposure' throughout the document; 4) Additional written instructions to address how to respond to signaling questions about temporality in a study of cross-sectional design a. For example, when responding to question 1.6 "Did authors avoid adjusting for post-intervention variable", we added "In a cross-sectional study, post-exposure variables are not studied and thus the action of adjusting or not adjusting for them does not present a risk to bias in the study. Therefore, the response option selected should represent that the risk to bias is not present or minimally present, not that the question is 'Not applicable'." 5) Additional instructions in conversations to address the subjectivity of the answer choices (for example the difference between 'Yes' and 'Probably Yes') and importance of explanations for why an answer choice was selected 6) Additional instructions in conversation to raters to minimize the use of the response option 'N/A'
Round 2: PFOA and fetal growth	1) Additional questions added to Domain 3. Bias in measurement of exposure to assess the exposure: a. "Is there a concern that the variation in exposure levels across groups was insufficient to potentially identify associations with health outcomes?"

<table>
<tr><td></td><td>b. “Is there a concern that the exposure assessment did not capture the relevant time window of exposure with respect to the health outcome?”
c. “Are there concerns that missing exposure data (including methods used to input data) may have resulted in exposure misclassification?”
2) Additional question added to Domain 3. Bias in measurement of exposure to assess temporality of exposure and outcome measurements:
a. “Was information on exposure status recorded prior to outcome assessment?”</td></tr>
<tr><td>Round 3: PBDE and thyroid function</td><td>1) Additional fields added to stage I of the instrument:
a. “List the criterial used to determine the accuracy of exposure measurement”
b. “List the possible co-exposures that could differ between exposure groups and could have an impact on study outcomes”
2) Additional fields added to stage II of the instrument:
a. “List the criteria used to determine the accuracy of exposure measurement”
b. “Factors to consider when evaluating health outcome assessment”</td></tr>
<tr><td>Consultation with topic-specific experts and ROBINS-I instrument developers</td><td>1) Discussions with topic-specific experts and comparison across instruments led to modifications made to the wording of questions in Domain 3: Bias in measurement of exposure and the inclusion of an additional question (3.7):
<table>
<tr><th>ROBINS-I (Bias in classification of intervention)</th><th>Modified instrument for assessing RoB in environmental exposure studies (Bias in measurement of exposure)</th></tr>
<tr><td>3.1 Is the intervention well defined?</td><td>3.1 Is exposure status well defined?</td></tr>
<tr><td>3.2 Was information on intervention status recorded at the time of intervention?</td><td>3.2 Did entry into the study begin with start of the exposure?</td></tr>
<tr><td>3.3 Was information on intervention status unaffected by knowledge of the outcome or risk of the outcome?</td><td>3.3 Was information on exposure status recorded prior to outcome assessment?</td></tr>
<tr><td></td><td>3.4 Could classification of exposure status have been affected by knowledge of the outcome or risk of the outcome?</td></tr>
<tr><td></td><td>3.5 Are the levels, duration, or range of exposure of the population at risk</td></tr>
</table></td></tr>
</table>

		sufficient or adequate to detect an effect of exposure?
		3.6 Is the follow-up period adequate to allow for the development of the outcome of interest?
		3.7 Were exposure assessment methods robust (including methods used to input data)?

2) Discussions with ROBINS-I instrument developers lead to the following modifications:

a. Reorganization of questions 3.5 and 3.6 into the project- and study-level protocols as measures to assess indirectness and study eligibility, not RoB
b. Agreement of replacing 'intervention' with 'exposure' throughout the instrument; replacement of 'target trial' with 'target experiment'; expansion of future guidance to distinguish between ROBINS for intentional interventions and modified ROBINS for unintentional exposures; and expansion of future guidance to highlight scenarios specific to environmental and occupational exposures.

Appendix D. Risk of Bias Instrument for Non-randomized Studies of Exposure.

Stage II: For each study

Specify a target experiment specific to the study:

☐ The protocol-specified target experiment fully applies ***OR***

Participant	
Experimental exposure	
Control exposure	

Specify the outcome

Specify which outcome is being assessed for risk of bias (typically from among those earmarked for the Summary of Findings table). Specify whether this is a proposed benefit or harm of exposure.

Is your aim for this study...?

- ☐ to assess the effect of initiating intervention (as in an intention-to-treat analysis)
- ☐ to assess the effect of initiating and adhering to intervention (as in a per-protocol analysis)
- ☐ other (specify)

Specify the numerical result being assessed

In case of multiple alternative analyses being presented, specify the numeric result (e.g. RR = 1.52 (95% CI 0.83 to 2.77) and/or a reference (e.g. to a table, figure or paragraph) that uniquely defines the result being assessed.

Preliminary consideration of confounders

Complete a row for each important confounding area (i) listed in the review protocol; and (ii) relevant to the setting of this particular study, or which the study authors identified as potentially important.

"Important" confounding areas are those for which, in the context of this study, adjustment is expected to lead to a clinically important change in the estimated effect of the exposure. "Validity" refers to whether the confounding variable or variables fully measure the area, while "reliability" refers to the precision of the measurement (more measurement error means less reliability).

(i) Confounding areas listed in the review protocol

Confounding area	Measured variable(s)	Is there evidence that controlling for this variable was unnecessary?*	Is the confounding area measured validly and reliably by this variable (or these variables)?	OPTIONAL: Is adjusting for this variable (alone) expected to move the effect estimate up or down?
			Yes / No / No information	Favor intervention / Favor control / No information

(ii) Additional confounding areas relevant to the setting of this particular study, or which the study authors identified as important

Confounding area	Measured variable(s)	Is there evidence that controlling for this variable was unnecessary?*	Is the confounding area measured validly and reliably by this variable (or these variables)?	OPTIONAL: Is adjusting for this variable (alone) expected to move the effect estimate up or down?
			Yes / No / No information	Favor intervention / Favor control / No information

* In the context of a particular study, variables can be demonstrated not to be confounders and so not included in the analysis: (a) if they are not predictive of the outcome; (b) if they are not predictive of exposure; or (c) because adjustment makes no or minimal difference to the estimated effect of the primary parameter. Note that "no statistically significant association" is not the same as "not predictive".

Preliminary consideration of criteria used to determine the accuracy of measurement of exposure and outcome

Complete a row for each measure listed in the study for the (i) exposure and (ii) outcome. Of the measures listed in the protocol, consider the sensitivity, specificity, and confidence in the methods used in the study.

(i) Exposure measurement method listed in the study		
Method of measurement	Measured exposure	Is the exposure measured validly and reliably by this method (or these methods)?
		Yes / No / No information

(ii) Outcome measurement method listed in the study		
Method of measurement	Measured outcome	Is the outcome measured validly and reliably by this method (or these methods)?
		Yes / No / No information

Preliminary consideration of co-exposures

Complete a row for each important co-intervention (i) listed in the review protocol; and (ii) relevant to the setting of this particular study, or which the study authors identified as important.
"Important" co-interventions are those for which, in the context of this study, adjustment is expected to lead to a clinically important change in the estimated effect of the intervention.

(i) Co-exposures listed in the review protocol		
Co-exposure	Is there evidence that controlling for this co-exposure was unnecessary (e.g., because it was not administered)?	Is presence of this co-exposure likely to favor outcomes in the experimental or the control group
		Favor experimental / Favor comparator / No information
		Favor experimental / Favor comparator / No information
		Favor experimental / Favor comparator / No information

(ii) Additional co-exposures relevant to the setting of this particular study, or which the study authors identified as important		
Co-exposure	Is there evidence that controlling for this co-exposure was unnecessary (e.g., because it was not administered)?	Is presence of this co-exposure likely to favor outcomes in the experimental or the control group
		Favor experimental / Favor comparator / No information
		Favor experimental / Favor comparator / No information
		Favor experimental / Favor comparator / No information

Stage III: For each study: risk of bias assessment

Risk of bias assessment (cohort-type studies)

Bias due to confounding	1.1 **Is there potential for confounding of the effect of exposure in this study? If N or PN to 1.1:** the study can be considered to be at low risk of bias due to confounding and no further signaling questions need be considered	Y / PY / PN / N	[Description]
	If Y/PY to 1.1, answer 1.2 and 1.3 to determine whether there is a need to assess time-varying confounding:		
	1.2. **<u>If Y or PY to 1.1</u>:** Was the analysis based on splitting follow up time according to exposure received? **If N or PN to 1.2**, answer questions 1.4 to 1.6, which relate to baseline confounding	NA / Y / PY / PN / N / NI	[Description]
	1.3. **If Y or PY to 1.2**: Were exposure discontinuations or switches likely to be related to factors that are prognostic for the outcome?	NA / Y / PY / PN / N / NI	[Description]
	If N or PN to 1.3, answer questions 1.4 to 1.6, which relate to baseline confounding		
	1.4. Did the authors use an appropriate analysis method that adjusted for all the critically important confounding areas?	NA / Y / PY / PN / N / NI	[Description]
	1.5. **If Y or PY to 1.4**: Were confounding areas that were adjusted for measured validly and reliably by the variables available in this study?	NA / Y / PY / PN / N / NI	[Description]
	1.6. Did the authors avoid adjusting for post-exposure variables?	NA / Y / PY / PN / N / NI	[Description]
	If Y or PY to 1.3, answer questions 1.7 and 1.8, which relate to time-varying confounding		

	1.7. Did the authors use an appropriate analysis method that adjusted for all the critically important confounding areas and for time-varying confounding?	NA / Y / PY / PN / N / NI	[Description]
	1.8. **If Y or PY to 1.7**: Were confounding areas that were adjusted for measured validly and reliably by the variables available in this study?	NA / Y / PY / PN / N / NI	[Description]
	Risk of bias judgement	Low / Moderate / Serious / Critical / NI	[Support for judgement]
	Optional: What is the predicted direction of bias due to confounding?	Favors experimental / Favors comparator / Unpredictable	[Rationale]
Bias in selection of participants into the study	2.1. Was selection of participants into the study (or into the analysis) based on variables measured after the start of the exposure? **<u>If N or PN to 2.1 go to 2.4</u>**	Y / PY / PN / N / NI	[Description]
	2.2. **<u>If Y/PY to 2.1:</u>** Were the post-exposure variables that influenced selection associated with exposure?	Y / PY / PN / N / NI	[Description]
	2.3. **<u>If Y/PY to 2.2:</u>** Were the post-exposure variables that influenced eligibility selection influenced by the outcome or a cause of the outcome?	NA / Y / PY / PN / N / NI	[Description]
	2.4 Do start of follow-up and start of exposure coincide for most participants?	NA / Y / PY / PN / N / NI	[Description]
	2.5 If Y/PY to 2.2 and 2.3, or N/PN to 2.4: Were adjustment techniques used that are likely to correct for the presence of selection biases?	NA / Y / PY / PN / N / NI	[Description]
	Risk of bias judgement	Low / Moderate / Serious / Critical / NI	[Support for judgement]
	Optional: What is the predicted direction of bias due to selection of participants into the study?	Favors experimental / Favors comparator / Towards null /Away from null / Unpredictable	[Rationale]

Bias in classification of exposures	3.1 Is exposure status well defined?	Y / PY / PN / N / NI	[Description]
	3.2 Did entry into the study begin with start of the exposure?	Y / PY / PN / N / NI	[Description]
	3.3 Was information used to define exposure status recorded prior to outcome assessment?	Y / PY / PN / N / NI	[Description]
	3.4 Could classification of exposure status have been affected by knowledge of the outcome or risk of the outcome?	Y / PY / PN / N / NI	[Description]
	3.5 Were exposure assessment methods robust (including methods used to input data)?	Y / PY / PN / N / NI	[Description]
	Risk of bias judgement	Low / Moderate / Serious / Critical / NI	[Support for judgement]
	Optional: What is the predicted direction of bias due to measurement of outcomes or exposures?	Favors experimental / Favors comparator / Towards null /Away from null / Unpredictable	[Rationale]
Bias due to departures from intended exposures	4.1. Is there concern that changes in exposure status occurred among participants? **If your aim for this study is to assess the effect of initiating and adhering to an exposure (as in a per-protocol analysis), answer questions 4.2 and 4.3, otherwise continue to 4.4 if Y or PY to 4.1.**	Y / PY / PN / N / NI	[Description]
	4.2. Did many participants switch to other exposures?	Y / PY / PN / N / NI	[Description]
	4.3. Were the critical co-exposures balanced across exposure groups?	Y / PY / PN / N / NI	[Description]
	4.4. **If NY/PN PY to 4.1, or Y/PY to 4.2, or 4.3:** Were adjustment techniques used that are likely to correct for these issues?	NA / Y / PY / PN / N / NI	[Description]
	Risk of bias judgement	Low / Moderate / Serious / Critical / NI	[Support for judgement]
	Optional: What is the predicted direction of bias due to departures from the intended exposures?	Favors experimental / Favors comparator / Towards null	[Rationale]

		/Away from null / Unpredictable	
Bias due to missing data	5.1 Were there missing outcome data?	Y / PY / PN / N / NI	[Description]
	5.2 Were participants excluded due to missing data on exposure status?	Y / PY / PN / N / NI	[Description]
	5.3 Were participants excluded due to missing data on other variables needed for the analysis?	Y / PY / PN / N / NI	[Description]
	5.4 **If Y/PY to 5.1, 5.2 or 5.3:** Are the proportion of participants and reasons for missing data similar across exposures?	NA / Y / PY / PN / N / NI	[Description]
	5.5 **If Y/PY to 5.1, 5.2 or 5.3:** Were appropriate statistical methods used to account for missing data?	NA / Y / PY / PN / N / NI	[Description]
	Risk of bias judgement	Low / Moderate / Serious / Critical / NI	[Support for judgement]
	Optional: What is the predicted direction of bias due to missing data?	Favors experimental / Favors comparator / Towards null /Away from null / Unpredictable	[Rationale]
Bias in measurement of outcomes	6.1 Could the outcome measure have been influenced by knowledge of the exposure received?	Y / PY / PN / N / NI	[Description]
	6.2 Was the outcome measure sensitive?	Y / PY / PN / N / NI	[Description]
	6.3 Were outcome assessors unaware of the exposure received by study participants?	Y / PY / PN / N / NI	[Description]
	6.4 Were the methods of outcome assessment comparable across exposure groups?	Y / PY / PN / N / NI	[Description]
	6.5 Were any systematic errors in measurement of the outcome unrelated to exposure received?	Y / PY / PN / N / NI	[Description]
	Risk of bias judgement	Low / Moderate / Serious / Critical / NI	[Support for judgement]

	Optional: What is the predicted direction of bias due to measurement of outcomes?	Favors experimental / Favors comparator / Towards null /Away from null / Unpredictable		[Rationale]
Bias in selection of the reported result	Is the reported effect estimate likely to be selected, on the basis of the results, from...?			
	7.1. ... multiple outcome *measurements* within the outcome domain?	Y / PY / PN / N / NI		[Description]
	7.2 ... multiple *analyses* of the exposure-outcome relationship?	Y / PY / PN / N / NI		[Description]
	7.3 ... different *subgroups*?	Y / PY / PN / N / NI		[Description]
	Risk of bias judgement	Low / Moderate / Serious / Critical / NI		[Support for judgement]
	Optional: What is the predicted direction of bias due to selection of the reported result?	Favors experimental / Favors comparator / Towards null /Away from null / Unpredictable		[Rationale]
Overall bias	**Risk of bias judgement**	Low / Moderate / Serious / Critical / NI		[Support for judgement]
	Optional: What is the overall predicted direction of bias for this outcome?	Favors experimental / Favors comparator / Towards null /Away from null / Unpredictable		[Rationale]

CHAPTER 4. RISK OF BIAS INSTRUMENT FOR NON-RANDOMIZED STUDIES OF EXPOSURES: A USERS' GUIDE

PREFACE TO CHAPTER 4

Chapter 4. *Risk of bias instrument for non-randomized studies of exposures: a users' guide* has been reviewed by all co-authors. This manuscript has been disseminated to members of the GRADE Environmental Health Project Group for review. The revised manuscript will be presented to the GRADE Working Group in April 2018 as a proposal for GRADE guidance. Following that meeting, the manuscript will be share among GRADE Working Group members and the GRADE Guidance Group for final approval. The final manuscript will be submitted to Environment International.

Risk of bias instrument for non-randomized studies of exposures: a users' guide

Author list

Rebecca L. Morgan [a]; Kristina A. Thayer [b]; Alison C. Holloway [c]; Nancy Santesso [a]; Robyn Blain [d]; Sorina Eftim [d]; Alexandra Goldstone [d]; Pam Ross [d]; Gordon Guyatt [a]; Elie A. Akl [a,e]; Mohammed T. Ansari [f]; Paul Whaley [g]; Holger J. Schünemann [a,h]

Affiliations

[a] Department of Health Research Methods, Evidence, and Impact, McMaster University, Health Sciences Centre, Room 2C14, 1280 Main Street West, Hamilton, ON L8S 4K1 Canada morganrl@mcmaster.ca, santesna@mcmaster.ca, guyatt@mcmaster.ca, ea32@aub.edu.lb, holger.schunemann@mcmaster.ca

[b] Integrated Risk Information System (IRIS) Division, National Center for Environmental Assessment (NCEA), Office of Research and Development, US Environmental Protection Agency, Building B (Room 211i), Research Triangle Park, NC USA 27711. thayer.kris@epa.gov

[c] Department of Obstetrics and Gynecology, McMaster University, Health Sciences Centre, Room 3N52A, 1280 Main Street West, Hamilton, ON L8S 4K1 Canada. hollow@mcmaster.ca

[d] ICF International Inc., 9300 Lee Highway, Fairfax, VA. Robyn.Blain@icfi.com, Pam.Ross@icfi.com, Ali.Goldstone@icfi.com, Sorina.Eftim@icfi.com

[e] Department of Internal Medicine, Faculty of Health Sciences, American University of Beirut, P.O. Box: 11-0236, Riad-El-Solh Beirut 1107 2020 Lebanon. ea32@aub.edu.lb

[f] School of Epidemiology and Public Health, Faculty of Medicine, University of Ottawa, ON K1H 8M5 Canada. tosansari@gmail.com

[g] Lancaster Environment Centre, Lancaster University, Lancaster LA1 4YQ, UK. p.whaley@lancaster.ac.uk

[h] Department of Medicine, McMaster University, Health Sciences Centre, Room 2C14, 1280 Main Street West, Hamilton, ON L8S 4K1 Canada. holger.schunemann@mcmaster.ca

Corresponding author: Holger J. Schünemann. Department of Clinical Epidemiology & Biostatistics, Health Sciences Centre, Room 2C14, 1280 Main Street West, Hamilton, ON L8S 4K1 Canada. holger.schunemann@mcmaster.ca

Conflict of interest

The authors declare they have no competing financial interests with respect to this manuscript, or its content, or subject matter.

Abstract

The objective of this paper is to explain how to apply, interpret, and present the results of a new instrument to assess the risk of bias (RoB) in non-randomized studies (NRS) dealing with effects of environmental exposures on health outcomes. This instrument is modeled on Risk Of Bias In Non-randomized Studies (ROBINS) of Interventions (ROBINS-I) instrument. The RoB instrument for NRS of exposures evaluates RoB along a standardized comparison to randomized experiments, instead of a study-design directed RoB approach. We provide specific guidance for the integral steps of developing a research question and target experiment, distinguishing issues of indirectness from RoB, making individual-study judgments, and performing and interpreting sensitivity analyses for RoB judgments across a body of evidence. To optimally integrate with an overall assessment of the certainty of evidence, we present an approach for integrating the RoB assessments in the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) approach. Finally, we guide the reader through an overall rating of all domains that determine the certainty of a body of evidence using the GRADE approach.

Abstract word count: 175 / 200

Keywords (6): Risk of bias; environmental health; GRADE; non-randomized studies; study limitations; ROBINS

Highlights

- To inform decision-making about an environmental health exposure, authors of systematic reviews should rigorously and transparently evaluate RoB of the included studies using a standardized approach.
- The RoB instrument for NRS of exposures presents a rigorous evaluation of RoB for individual studies for each outcome.
- At the review level, overall RoB across the body of evidence for a specific outcome is a crucial part of judging the certainty of that evidence when using the GRADE approach.

Abbreviations

BMI: body mass index

BPA: bisphenol-A

CoE: Certainty of evidence

dB: decibel

GRADE: Grading of Recommendations Assessment, Development, and Evaluation

NHANES: National Health and Nutrition Examination Survey

NRS: Non-randomized studies

PECO: population, exposure, comparator, outcome

RCT: randomized controlled trial

RoB: Risk of bias

ROBINS-E: Risk of Bias in Non-randomized Studies of Exposures

ROBINS-I: Risk of Bias in Non-randomized Studies of Interventions

uBPA: urinary measure of bisphenol-A

1. Introduction

The evidence on the impact of environmental exposure on health outcomes typically comes from non-randomized studies (NRS). Objective and transparent evaluation of evidence requires the use of systematic reviews [1]. A highly credible systematic review requires a standardized, rigorous, and transparent evaluation of the risk of bias (RoB) in each included study across the body of evidence [2, 3].

A recent study evaluated five RoB methods used in environmental health hazard assessments [4]. While all five methods considered the same issues (or domains) in RoB assessment, their relative emphasis on these issues varied. These findings demonstrated the need for the harmonization and improvement of these methods.

The objective of this paper is to explain how to apply, interpret, and present the results of a new instrument to assess the RoB in NRS dealing with effects of environmental exposures on health outcomes.

2. Overview of the instrument

The RoB instrument for NRS of exposures is modeled after the Risk Of Bias In Non-randomized Studies (ROBINS) for interventions (ROBINS-I) instrument. It uses an ideal target experiment as a reference point. Hernan et al. proposed that causal inference from NRS represent an attempt to emulate the ideal randomized experiment (the target experiment) that would answer the question of interest [5]. By using a the target experiment as the reference point, ROBINS moves away from a study-design directed

approach, i.e., using the specific design of the observational study as part of the RoB assessment [6].

In brief, the proposed RoB instrument for NRS of exposures is completed in three stages:

1. Stage I: present the review question, confounders, co-interventions, and exposure and outcome measurement accuracy information;
2. Stage II: describe each eligible study in relation to a hypothetical target experiment, including confounders and co-interventions that will require consideration; and
3. Stage III, assess RoB across seven domains about the strengths and limitations of studies of environmental exposure.

To distinguish between the term domain employed by the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) framework for the assessment of certainty of a particular effect estimate across a body of evidence, of which the overall RoB is one, we refer to the instrument's domains from here on as 'items'. The seven RoB items are: 1) bias due to confounding, 2) bias in selection of participants into the study, 3) bias in classification of exposures, 4) bias due to departures from intended exposures, 5) bias due to missing data, 6) bias in measurement of outcomes, and 7) bias in selection of reported results. Judgments for each RoB item can be: 'Low RoB', 'Moderate RoB', 'Serious RoB', and 'Critical RoB'. In order to reach a judgment for each RoB item, the rater answers first one or more signaling questions with 'Yes', 'Probably yes', 'Probably

no', or 'No'. The answer should be based on the information available in the individual study and be justified in an accompanying free-text field. Similarly, an overall judgment about the bias in an individual study is either 'Low RoB', 'Moderate RoB', 'Serious RoB', and 'Critical RoB'.

Previously published guidance for the ROBINS-I instrument proposes that the study-level RoB should be the most concerning level among the RoB items for that study, unless raters determine the study-level RoB to be more severe because of compounded risks than an individual RoB item [7]. Identifying RoB per item and per individual study allows systematic-review authors to explore the possible influence of lower versus higher RoB studies on the pooled estimates of effect from a synthesis of studies [8]. Of note is that the study-level risk of bias is assessed for each outcome in a study, as such, study RoB could vary by the outcomes (e.g. subjective outcomes may have different biases than for objective outcomes). Therefore, the RoB instrument for NRS of exposures would need to be completed for all relevant outcomes.

Systematic-review authors can then use the RoB instrument as part of the assessment of the certainty of the body of evidence using the GRADE framework. Within the GRADE framework, RoB is one domain for assessing the certainty of evidence (CoE), the others being inconsistency, indirectness, imprecision, publication bias, magnitude of effect, dose-response gradient, and plausible opposing residual confounding [2]. When assessing the CoE from NRS, the evidence previously entered GRADE with an initial certainty of 'Low'. However, since the RoB instrument for NRS of exposures takes into account lack of randomization, evidence would not be automatically rated down.

Therefore, bodies of evidence of any study design will undergo a detailed RoB evaluation without the influence of study-design labels. All included studies within the bodies of evidence will start at the same 'High' initial certainty within GRADE. NRS will typically be more likely to be rated down for RoB based on increase potential for bias when compared with randomized controlled trials (RCTs) [6].

When conducting a systematic review, results from the study-level RoB instrument for NRS of exposures assessment can be synthesized to inform judgments about overall RoB; however, guidance on the application of the RoB instrument for NRS of exposures does not exist. This article provides initial guidance and procedural steps for the application of the RoB instrument for NRS of exposures to individual studies and across a body of evidence to reach an overall RoB judgment in the GRADE framework [9]. Although the RoB instrument for NRS of exposures is still being refined in consultation with a diverse group of subject matter experts, we highlight a number of important procedural questions that have been identified during the course of developing this version. Thus, dissemination of our experience in implementing a RoB instrument for NRS of exposures should facilitate future testing and clarify intended usage of the tool.

3. Approach when conducting systematic reviews for studies of exposure

We previously described the development of the RoB instrument for NRS of exposures [10]. In addition to this effort, we have solicited broader input on this instrument at workshops held at GRADE Working Group meetings in March 2015, October 2015, and

April 2016; during a meeting to develop ROBINS of Exposures (ROBINS-E; an instrument based on the RoB instrument for NRS of exposures and ROBINS-I) in January 2017; and at the Global Evidence Summit in September 2017. These workshops have led to further refinement and pilot-testing of ROBINS for exposure.

Figure 1 presents a schematic of how the RoB instrument for NRS of exposures instrument fits into the systematic-review process. It illustrates steps for evaluating the RoB of individual studies in a review and integrating the results across a body of evidence into the GRADE evidence-assessment framework. For each outcome in the review, authors of systematic reviews would go through Stages II and III, and GRADE.

3.1. Complete Stage I of the RoB instrument for NRS of exposures

3.1.1. Define the research question

This process begins with the clarification of a research question. For questions about unintentional interventions, i.e., exposures, namely the environmental and occupational type, the research question is formatted as a PECO (population, exposure, comparator(s), and outcomes) question [10, 11]. We demonstrate this in the following example about noise-level exposure and hearing loss. Understanding the relationship between decibel (dB) level exposure and hearing loss informs the PECO. “In shift workers, what is the effect of 80 dB of sound intensity or greater compared to less than 80 dB on hearing loss and other outcomes?”, where the exposure is ≥ 80 dB and the comparison is < 80 dB.

Because the RoB instrument for NRS of exposures is set up as a comparison between groups that can be exposed or not, or with different levels of exposure, it is necessary to clearly identify what is the exposure and what is the comparison. In some situations, not all information needed to formulate a PECO is available. There are at least five paradigmatic scenarios to facilitate formulating and defining the "E" and the "C" within the PECO: 1) select the comparator based on what exposure cut-offs (e.g., thresholds, levels, durations, ranges, means, medians, or ranges of exposure) can be achieved through an intervention; 2) use existing exposure cut-offs associated with the known health outcomes of interest; 3) when only the exposure for a population is known, use cut-offs from external or other populations (may come from other research); 4) use cut-offs defined based on the distribution in studies identified in the review; and 5) explore the shape and distribution of the relationship between the exposure and outcome (e.g., risk of the health outcome from an incremental increase in the exposure) [12]. Researchers should be transparent about which of these recommendations they are addressing with their PECO and ensure that the exposure and comparator(s) are explicitly defined.

Our paradigmatic PECO scenarios suggest that when the pervasiveness and unintentional nature of the exposure make 'no exposure' unlikely or unfeasible, the comparator will be a different level of exposure [12]. For example, if presented with a policy request to determine at what cut-off should occupational noise exposure be limited to in order to prevent hearing loss among shift workers, the sensible comparison may be an alternative dB level, since a comparison of 'no noise' is unachievable. The

objective is to determine a comparison where the rater will be able to distinguish between persons who have received different levels of exposure (e.g., more exposure than the minimum). Raters and topic-specific experts should take into consideration how willing they are to accept certain cut-offs used to differentiate between compared groups. If cut-offs are not known, one of the other scenarios for the PECO will apply.

The example of determining the relationship between dB exposure and hearing loss for persons exposed to less than 20 dB introduces an additional concept of "study sensitivity" (i.e., the ability of a study to detect a true effect or hazard) [13]. In the situation of dB exposure, the spread/range of the exposure and a health outcome (hearing loss) is known to be broad; therefore, we understand that comparing the effects of dB levels between 1 and 20 will fail to demonstrate a difference in hearing loss that would exist for higher levels or wider intervals. However, if we compare ≥ 80 dB to < 80 dB, we would have greater confidence in the ability to detect a true effect. For rare exposures or exposures with less information about the level, duration, frequency, or probability of exposure, the inability to detect a true effect cannot be excluded if the exposure levels are narrow.

3.1.2. Identify confounders, co-interventions, and measures of exposures and outcomes

In Stage I, review authors complete free-text fields to list confounders and co-interventions that are associated with both the exposure and outcome. In addition, review authors complete fields on the accuracy of the exposure and outcome measurements. These sections can be populated by any knowledgeable member of the

review author team. Working through these sections, raters respond to signaling questions in the confounding, participant selection, and exposure measurement items. Consideration of these issues may also illuminate when sources of indirectness may occur [10]. For example, the review team may identify that one of their outcomes is body weight; however, studies may measure waist circumference (and measure it accurately within the study) to inform the outcome of body weight. The review team may determine waist circumference to be an indirect measure of weight.

We present the text used in the review-level protocol for an example on bisphenol-A (BPA; comparing highest and lowest levels of exposure to BPA) in Appendix A. The PECO being: "What is the effect of highest levels vs. lowest levels of BPA exposure on body weight?" We reviewed published literature, as well as consulted with topic-specific experts, to determine the final set of responses to the Stage I fields. For some exposures, a public database of confounders for measures of environmental exposures and health outcomes (i.e., PhenX Toolkit; https://www.phenxtoolkit.org/) may provide additional information.

3.2. Complete Stage II of the RoB instrument for NRS of exposures for presumably eligible studies

3.2.1. Construct the target experiment

At this point, the studies that meet the eligibility criteria of the review have been identified. The RoB instrument for NRS of exposures will need to be completed for each relevant outcome within each study. In Stage II, reviewers construct a study-specific target experiment (i.e., ideal hypothetical RCT) to mimic it, by specifying the exposure

and comparator in the study, exposure thresholds, outcome-specific confounders, and health outcome measurements. As explained in previous GRADE guidance for the use of ROBINS-I, establishing the target experiment provides a structured comparison (i.e., along an absolute scale) with a reference study that is presumably at the lowest RoB [6]. It then allows RoB assessment of individual studies and across studies at a later stage against the lowest possible bias that research could yield for the question at hand. Also in Stage II, the reviewer records how the individual studies measured the exposure and health outcome. The information recorded in Stage II informs the RoB judgments made in Stage III.

For example, we consider our review on BPA and weight. The PECO of the review compares a higher to a lower level of BPA exposure. In Stage II, we determine the target experiment for an included study (Appendix B). Based on the study by Carwile & Michels, the target experiment would be framed as: "In the general adult population, what is the effect of exposure to BPA highest levels (quartile 4: ≥ 4.7 ng/mL) compared to BPA lowest levels (quartile 1: ≤ 1.1 ng/mL) on body weight?" In this situation, we compared two exposure cut-offs to determine the effect on the outcome of body weight, as measured within the study as obesity.

Confounders must be explored in each individual study determined eligible for the review, as each study may introduce different confounders (e.g., the review question may be about the general population, but the study includes only industrial workers which may introduce additional confounders, such as exposure to other chemicals – note that it may have impact on judging indirectness, too). In Stage II, the reviewer

makes a judgment as to the potential magnitude and direction of the impact of the confounding on the effect estimate. For example, when examining the effect of BPA on body weight, consumption of processed foods is considered a confounder as it both increases the participants' exposure to BPA through food packaging and increases overall caloric intake [14]. We present the completed Stage II sections for two studies from our BPA and obesity example: Carwile & Michels, 2011 and Harley et al., 2013 (Appendices B & C) [15, 16].

3.2.2. Identifying sources of indirectness to integrate within GRADE and their relation to risk of bias

While establishing the target experiment in Stage II, individuals may identify studies that present evidence different from the PECO question (i.e., a restricted version of any concept) [17]. For example, consider again the review of hearing loss due to noise exposure. Studies with only shift workers may be considered indirect evidence for effects in the general population. Studies reporting on waist circumference may be considered indirect evidence for the measure of the exposure. Sources of indirectness may also come from studies that do not have a comparison (and therefore results would be compared to an external control or comparator) or when using surrogate measures. While the review team may decide to include this study in the review, when evaluating the evidence within GRADE differentiation between the domain of risk of bias and indirectness may be rather nuanced.

When formulating the target experiments in Stage II, for some reviews, authors may decide that studies that do not make a comparison could be included (i.e., only the

effects of the exposure are provided, such as in a case report or case series). In such studies, an indirect comparison, such as a historical control (e.g., lung-cancer mortality before the introduction of coal-fueled factories) of a population without that exposure from another study could be used as the comparison. Studies for which there is an indirect control have higher RoB given the expected imbalance between prognostic factors that predict outcomes (i.e., confounders).

Subsequent considerations about RoB when using indirect evidence in a review require critical evaluation of misclassification of the exposure. While it is important to recognize the potential for more serious bias in classification of exposure when using an indirect comparison, there are situations in which they may present less risk because of clearly delineated exposure and comparison groups (if the exposure measurement methods are reliable and valid). For example, one study potentially eligible for inclusion in the BPA and body weight review presents a single mean estimate of BPA (2.27 ng/mL ± 0.32 ng/mL) [18]. This aggregated exposure measurement prevents an in-study comparison of highest vs. lowest levels of BPA. Since the study does not provide disaggregated data for a comparator group, one option is to identify a comparator from an external source, such as a historical control presenting a BPA and body weight group outside of the values in the eligible study (e.g. an estimate lower than 2.27 ng/mL). If the comparator BPA levels are exclusive of those in the eligible study, most likely misclassification of the exposure levels will not be a concern provided pre-analytic and analytical validity of respective BPA assays in terms of sampling, storage conditions, and assay properties are judged to be comparable between the two sources of evidence. If there is uncertainty

about whether or not the range of BPA exposure levels overlaps, then misclassification of the exposure may be more of a concern. Also, the instruments used to measure the exposure can inform the risk of misclassification, which may introduce greater bias as they may be less comparable.

Similarly, studies identified for the review may use exposure measures indirect to those identified in the PECO, i.e., surrogate measures. Within the BPA example, we were interested in the amount of BPA consumed; however, an oral measure of BPA was not obtainable. Therefore, we accepted studies presenting a urinary measure of BPA (uBPA). Further exploration of the literature and consultation with topic-matter experts confirmed uBPA as a direct, reliable, and accurate measure of oral BPA; however, the measurement of BPA exposure level based on a participant's job title (e.g. cashier) would be indirect [19]. Extrapolating BPA exposure levels based on a participant's job title may also introduce a risk to bias based on specific prognostic factors or the ability to differentiate between the levels of exposure.

3.3. Complete Stage III of the RoB instrument for NRS of exposures assessment for eligible studies

Raters evaluate eligible studies and determine RoB by responding to signaling questions across the seven RoB items listed previously. Appendices D & E present summaries from two studies addressing BPA and body weight (as measured by prevalent overweight and prevalent obesity). We present judgments across the RoB instrument items for NRS of exposures in a RoB matrix for all eligible studies in Table 1.

Due to the lack of randomization and unintentional nature of the exposure, studies will typically be judged as 'Serious' RoB within the item of bias due to confounding, but also may often be judged as 'Serious' due to selection of participants, and measurement of exposure. While RoB items 4-to-7 are similar to those used to evaluate RCTs [7, 20], bias due to confounding, selection of participants, and classification of the exposure present considerations unique to studies of exposures [10]. Below, we highlight some of these nuances and how raters can address them in their item and study-level RoB judgments.

3.3.1. Bias due to Confounding

Three situations are common when evaluating bias due to confounding for environmental exposures: 1) the evaluation of cross-sectional studies; 2) considerations of large effect or opposing residual confounding, and 3) assessing the impact of unmeasured confounding.

When considering bias due to confounding, cross-sectional studies present a situation that can impact the item-level RoB judgment. This is because we are unable to evaluate time-varying confounding and it makes the measurement of the effect of known confounders more difficult. We present two examples from the BPA and weight review. While Carwile & Michels adjusted for all critical confounders, the measurement of exposure and outcome at one time point lowers our certainty that confounders (e.g. dietary preference for canned food) are not responsible for any observed association (Appendix D) [15]. In this specific study, the data collection point is part of the National Health and Nutrition Examination Survey (NHANES), a nationally-representative dataset with years of prior data collection, therefore providing supplemental information about

the adjustment of confounders. In contrast, within that review, neither Li nor Wang provide that same level of information about the data collection, therefore presenting "Critical" bias due to confounding (Table 1) [21, 22].

Studies judged as biased due to confounding with evidence of a large effect or opposing residual confounding may not require the most severe RoB item-level judgment. This is due to the magnitude of the effect outweighing the size of the bias that we have observed or that all plausible biases go in a direction that would have reduced the observed effect or increased the observed lack of effect. These two factors are typically considered within the GRADE evidence assessment for NRS as a method of increasing the CoE; however, within the RoB instrument for NRS of exposures they may also influence the study-level judgments [23]. To demonstrate this situation, we present an example on smoking and lung cancer-related mortality [24, 25]. A prospective cohort study compared lung cancer-related mortality rates among smokers and non-smokers [25]. Although there are some concerns with confounding due to residual and unmeasured confounders, such as occupational or air pollution exposures, the large magnitude of effect (30 times greater mortality rate due to lung cancer among persons smoking 25 or more cigarettes vs. non-smokers) warrants a less severe RoB item-level judgment of 'Low' or 'Moderate', instead of 'Serious' for the RoB item of confounding [25]. In this example, the large magnitude of effect reduces our concern that the effect is spurious [26]. This is how the concepts of magnitude of effect and opposing residual confounding would be incorporated into the individual study-level RoB instrument for NRS of exposures assessment. In addition, exploratory research conducted since the

time of publication has suggested no correlation between occupational exposures to the 10 most common exposures (i.e., sulfur dioxide, welding fumes, engine emissions, gasoline, lubricating oil, solvents, paints/varnishes, adhesives, excavation dust, and wood dust) and smoking history [27]. This exploration into the relationship between exposures and our outcome of interest reduces our concern for potential residual and unmeasured confounding due to other occupational or air pollution exposures even more.

3.3.2. Bias due to Misclassification of Exposure

In studies of exposure, there is a particular concern with distinguishing between the exposed and reference groups, as measuring exposure is difficult. To continue with the example of BPA and body weight, the review team and topic-specific experts note the accuracy of the measurement of exposure requires multiple measurements (cited here from five-to-13 repeated measurements) at different time points, due to the non-persistent nature of BPA in the body [28]. If an individual study uses fewer than the recommended number of samples, or since diagnostic accuracy of BPA with the collection of between five and 13 samples only yields ≥ 0.80 sensitivity and specificity depending on level of exposure (small, moderate, high), there are concerns for non-differential misclassification (i.e. random error) potentially conflating participants in the exposure and comparator groups, likely leading to little difference in the outcomes (i.e. bias toward the null). When considering two studies measuring exposure to BPA, for Harley, there is less of a concern about non-differential misclassification because at least four samples were collected from each participant (Appendix E) [16]. While the number

of samples increases our certainty in the correct classification of the higher exposed and lower exposed groups, the number is still not enough and we still have serious concerns about the correct classification of the exposure groups in some studies. In Carwile & Michels, participants provided only one sample (Appendix D) [15]. The single sampling method used in Carwile & Michels decreases our certainty that the higher exposed and lower exposed participants can be accurately distinguished. Both studies introduce risk due to the misclassification of the exposure levels because neither meet the minimum recommended required number of samples. Returning to figure 1, in their protocol, review authors could have specified to exclude such studies a priori or identified this risk of bias item as a reason to conduct a sensitivity analysis (see below).

3.4. RoB judgments for an individual study for an outcome

Per study, the most critical of the RoB item-level judgments determines the study-level RoB, unless raters determined the study to have more severe RoB based on a combination of RoB judgments across items. We demonstrate this with our example of BPA and weight in Table 2. This approach relies on individuals critically evaluating the rationale and direction of the bias. For example, if more than one RoB items within a study were rated as serious RoB but no RoB items were of critical RoB, then the study-level RoB could either be serious or could be critical if the consideration of all serious ratings leads to greater concern than would be expressed by a rating of serious on the study level. Raters should justify this assessment.

3.5. Sensitivity analyses and overall RoB across studies

Sensitivity analyses allow for the exploration of heterogeneity across a body of evidence to determine if there are concerns with including studies with certain RoB [29]. The objective is to identify different sources of bias that might influence variability in the effect estimate. The variability introduced by RoB items may inform whether a subgroup of studies, rather than the whole body of evidence, best informs the research question. The approach to conducting sensitivity analyses (not to be confused with the sensitivity of a study) should be specified at the protocol stage of the systematic review. For example, studies deemed critical in the domain of Bias due to confounding may result from unadjusted analyses of covariates. If a body of evidence includes studies with adjusted and unadjusted analyses, a sensitivity analysis could compare the estimates of effect for the adjusted (removing those studies not adjusting for covariates) and the total pooled estimate. If the effect estimates are not robust and differ between analyses (e.g. confounding may have an influence on the results), then review authors might consider whether to exclude the studies with unadjusted analyses. However, if the effect estimates do not differ (e.g. confounding has no influence on the results), then the unadjusted studies may remain in the analysis because the suspicion of confounding is not corroborated. In these instances when the effect estimate is similar across studies then authors could consider updating the individual study level ratings to indicate low risk of bias for the item and include the rationale that the sensitivity analysis showed no effect of the risk of bias on the results.

Using BPA as an example, we compared studies for the outcomes of prevalent overweight and prevalent obesity at higher and lower RoB in sensitivity analyses

(Appendices F & G). The sensitivity analysis for the outcome of prevalent overweight resulted in a difference between the effect estimates, demonstrating that bias due to confounding impacted the pooled estimate; therefore, the judgment would be reflective of the more severe RoB (Table 3). An additional option would be to only show results from Harley, Eng, and Carwile in the GRADE evidence assessment. In contrast, the sensitivity analysis of studies reporting on prevalent obesity demonstrated similar effect estimates (Appendix G). In this situation, all studies reflect the less serious RoB judgment (Table 4).

3.6. Integration of RoB judgment across a body of evidence into GRADE assessment

The overall rating of RoB across the body of evidence for an outcome is integrated into the GRADE assessment similar to what has been previously described for the result of RCTs and observational studies [8]. In addition, we described situations where indirectness may be captured in Stages I or II within the RoB instrument for NRS of exposures, but would be integrated in the overall assessment of the evidence. When evaluating RoB using ROBINS-I and the RoB instrument for NRS of exposures, studies start at 'High' initial CoE within GRADE. For the example of BPA and its effect on weight, we present the outcomes of prevalent overweight (i.e., body mass index [BMI] ≥85th percentile for age/gender in children; BMI 18.5-25/30 kg/m2) and prevalent obesity (BMI ≥95th percentile for age/gender in children; BMI ≥25-30 kg/m2) in a GRADE evidence profile (Table 5). It is across this body of evidence that we look for evidence of the three factors (magnitude of effect, dose-response gradient, and opposing residual

confounding) considered in the past as mechanisms to upgrade the quality of the evidence for NRS within GRADE [23]. The BPA example does not demonstrate any situation which may lead to a less severe RoB judgment. Across the body of evidence for prevalent overweight, our RoB based on the RoB instrument for NRS of exposures evaluation and sensitivity analysis is 'Critical', necessitating rating down three levels for RoB. In addition, we rate down for imprecision because the effect includes both benefit and harm. Our final certainty for this outcome would be 'Very low'. Across the body of evidence for prevalent obesity, our RoB is 'Serious'; therefore, we rate down two levels for RoB. There are no other GRADE domains that we would rate down for. Our final CoE would be 'Low'.

4. Discussion

The RoB instrument for NRS of exposures presents a novel, standardized and structured instrument for assessing individual studies and across studies in a systematic review. We present an overview of the process, using examples to demonstrate specific issues encountered when formulating the PECO for the review, outlining a target experiment for an individual study, evaluating bias in individual studies, and summarizing judgments across the body of evidence. We highlight the need for critical appraisal of the RoB judgments, including situations within individual studies and across a body of evidence when the judgments may be less severe. In addition, we present sources of indirectness identified in eligible studies that would inform the GRADE evidence assessment. We also

present the steps for integrating the RoB across a body of evidence into a GRADE evidence profile.

4.1. Advantages and disadvantages of using the RoB instrument for NRS of exposures approach

Some challenges remain, specifically when defining the target experiment and making judgments at the study and body of evidence level. The major limitation of identifying a target experiment question is that much of the research on environmental health exposures is still trying to assess the potential link with a human health hazard. Defining a specific comparison to an exposure presents a challenge, as there may be a paucity of evidence to support the distinct exposure and comparator. However, we utilize our work on defining appropriate questions and present five scenarios to facilitate the identification of an exposure and comparator [12]. In addition, the best available studies to inform a review may only present data on one exposure category. In this situation, raters can draw on sources of comparative exposure data, such as historical controls, ideally summarized in a systematic review.

Interrater reliability has not yet been measured, however, the purpose of the RoB instrument for NRS of exposures is not to reach the same judgment and across studies, but instead to justify the judgements and make the judgements transparent. We present several examples when using the RoB instrument for NRS of exposures; however, more examples are needed to highlight nuances of this instrument when applied by and across studies.

Based on concerns from systematic-review authors and guideline developers in the environmental health field, the RoB instrument for NRS of exposures is the first to evaluate bias using a standardized comparison to a target experiment. Although the risk of bias in individual and across studies will not change, this approach allows the body of evidence to start at 'High' initial CoE in the GRADE framework, potentially improving acceptability of this instrument and the use of GRADE for environmental decision-making assessments.

4.2. Relation to other studies

This is the first article describing examples from systematic reviews using the RoB instrument for NRS of exposures to evaluate the RoB across a body of evidence for a specific outcome. We present one option of a RoB matrix displaying the RoB study- and item-level judgments. In addition, we present examples of when an individual and a body of evidence RoB judgment may be improved (determined to be a less severe RoB) based on further exploration of residual and unmeasured confounding. We highlight the value added by performing sensitivity analyses with the body of evidence to explore sources of bias.

The application of ROBINS-I for RoB assessment across a body of evidence is undergoing further development, as are the procedures for interpreting RoB within the GRADE approach when NRS are compared to RCTs as in the RoB instrument for NRS of exposures or ROBINS-I [6]. Collaboration between the developers of the RoB instrument for NRS of exposures and these projects allows for an iterative approach to methods advancements.

4.3. Implications for stakeholders using the RoB instrument for NRS of exposures

Evaluating the RoB across the body of evidence for an outcome informs one domain within the GRADE framework's evidence assessment contributing to the understanding about the overall CoE. The RoB instrument for NRS of exposures allows systematic-review authors and guideline developers the opportunity to start at 'High' initial CoE within GRADE. Using this instrument should not result in a final certainty distinct from the prior approach of starting NRS at 'Low' initial CoE within GRADE because the conceptual underpinnings are the same. However, the approach does not rely on often poorly defined study labels and is more transparent. Indeed, users may prefer investigating the relationship between rating down for imbalances due to confounders, selection bias, or misclassification of the exposure instead of starting at 'Low' initial CoE as a general judgment about these items. The process and examples outlined in this manuscript provide guidance for researchers and guideline developers using evidence about exposures to inform their systematic reviews and decision making.

4.4. Unanswered questions and future research

While we present situations of where magnitude of effect and opposing residual confounding may decrease our concerns about bias within both individual assessments and across the body of evidence, more exploration of the role of dose-response is needed. Future research should provide examples of how to incorporate dose response into an assessment using the RoB instrument for NRS of exposures.

5. Conclusions

The RoB instrument for NRS of exposures provides a novel approach for evaluating RoB of exposures. Determining the RoB across a body of evidence is critical to inform decision making about environmental health exposures. We present guidance and examples for systematic-review authors and guideline developers to follow when using this instrument.

6. Acknowledgments

We are grateful to all systematic review and GRADE members who have collaborated with their feedback and suggestions for this work.

7. Funding Sources

This research was supported by the Intramural Research Program of the National Institute of Environmental Health Sciences and the MacGRADE Centre at the McMaster University.

8. References

1. Woodruff TJ, Sutton P: **The Navigation Guide systematic review methodology: a rigorous and transparent method for translating environmental health science into better health outcomes**. *Environ Health Perspect* 2014, **122**(10):1007-1014.
2. Balshem H, Helfand M, Schunemann HJ, Oxman AD, Kunz R, Brozek J, Vist GE, Falck-Ytter Y, Meerpohl J, Norris S *et al*: **GRADE guidelines: 3. Rating the quality of evidence**. *J Clin Epidemiol* 2011, **64**(4):401-406.
3. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JP, Clarke M, Devereaux PJ, Kleijnen J, Moher D: **The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration**. *PLoS medicine* 2009, **6**(7):e1000100.
4. Rooney AA, Cooper GS, Jahnke GD, Lam J, Morgan RL, Boyles AL, Ratcliffe JM, Kraft AD, Schünemann HJ, Schwingl P: **How credible are the study results? Evaluating and applying internal validity tools to literature-based assessments of environmental health hazards**. *Environment international* 2016.
5. Hernán MA, Robins JM: **Using big data to emulate a target trial when a randomized trial is not available**. *American journal of epidemiology* 2016, **183**(8):758-764.
6. Schünemann H, Cuello C, Akl EA, Mustafa R, Meerpohl J, Thayer K, Morgan R, Gartlehner G, Kunz R, Katikireddi S *et al*: **GRADE Guidelines: 18. How tools to assess risk of bias in non-randomized studies should be used to rate the certainty of a body of evidence**. Unpublished.
7. Sterne JA, Hernan MA, Reeves BC, Savovic J, Berkman ND, Viswanathan M, Henry D, Altman DG, Ansari MT, Boutron I *et al*: **ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions**. *BMJ* 2016, **355**:i4919.
8. Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, Montori V, Akl EA, Djulbegovic B, Falck-Ytter Y *et al*: **GRADE guidelines: 4. Rating the quality of evidence--study limitations (risk of bias)**. *J Clin Epidemiol* 2011, **64**(4):407-415.
9. Morgan R, Thayer K, Holloway A, Santesso N, Blain R, Eftim S, Goldstone A, Ross P, Guyatt G, Schünemann H: **Application of a Risk of Bias in Non-randomized Studies of Exposure (ROBINS for exposures) instrument and integration into certainty in the evidence assessments using GRADE**. Unpublished.
10. Morgan RL, Thayer K, Santesso N, Holloway AC, Blain R, Eftim S, Goldstone A, Ross P, Guyatt G, Schünemann H: **Need for an instrument to evaluate Risk of Bias in Non-randomized Studies of Exposure: Rationale and preliminary instrument**. *Environment International* Under review.
11. Morgan RL, Thayer KA, Bero L, Bruce N, Falck-Ytter Y, Ghersi D, Guyatt G, Hooijmans C, Langendam M, Mandrioli D *et al*: **GRADE: Assessing the quality of evidence in environmental and occupational health**. *Environ Int* 2016, **92-93**:611-616.

12. Morgan RL, Thayer K, Whaley P, Schünemann H: **Identifying the PECO: A framework for formulating good questions to explore the association of environmental and other exposures with health outcomes**. Unpublished.
13. Cooper GS, Lunn RM, Agerstrand M, Glenn BS, Kraft AD, Luke AM, Ratcliffe JM: **Study sensitivity: Evaluating the ability to detect effects in systematic reviews of chemical exposures**. *Environ Int* 2016, **92-93**:605-610.
14. Ranciere F, Lyons JG, Loh VH, Botton J, Galloway T, Wang T, Shaw JE, Magliano DJ: **Bisphenol A and the risk of cardiometabolic disorders: a systematic review with meta-analysis of the epidemiological evidence**. *Environ Health* 2015, **14**(1):46.
15. Carwile JL, Michels KB: **Urinary bisphenol A and obesity: NHANES 2003-2006**. *Environ Res* 2011, **111**(6):825-830.
16. Harley KG, Schall RA, Chevrier J, Tyler K, Aguirre H, Bradman A, Holland NT, Lustig RH, Calafat AM, Eskenazi B: **Prenatal and postnatal bisphenol A exposure and body mass index in childhood in the CHAMACOS cohort**. *Environmental health perspectives* 2013, **121**(4):514.
17. Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, Alonso-Coello P, Falck-Ytter Y, Jaeschke R, Vist G *et al*: **GRADE guidelines: 8. Rating the quality of evidence--indirectness**. *Journal of clinical epidemiology* 2011, **64**(12):1303-1310.
18. Zhao HY, Bi YF, Ma LY, Zhao L, Wang TG, Zhang LZ, Tao B, Sun LH, Zhao YJ, Wang WQ *et al*: **The effects of bisphenol A (BPA) exposure on fat mass and serum leptin concentrations have no impact on bone mineral densities in non-obese premenopausal women**. *Clinical Biochemistry* 2012, **45**(18):1602-1606.
19. Thayer K, Rooney A, Boyles A, Holmgren S, Walker V, Kissling G, U.S. Department of Health and Human Services: **Draft protocol for systematic review to evaluate the evidence for an association between bisphenol A (BPA) exposure and obesity**. *National Toxicology Program* 2013.
20. Higgins J, Sterne J, Savovic J, Page M, Hrobjartsson A, Boutron I, Reeves B, Eldridge S: **A revised tool for assessing risk of bias in randomized trials** In: *Cochrane Methods.* Edited by Chandler J, McKenzie J, Boutron I, Welch V. http://www.cochranelibrary.com/dotAsset/ecafc5c7-0b9b-4cd1-a4c1-8b0013aea046.pdf; 2016.
21. Li D-K, Miao M, Zhou Z, Wu C, Shi H, Liu X, Wang S, Yuan W: **Urine bisphenol-A level in relation to obesity and overweight in school-age children**. *PloS one* 2013, **8**(6):e65399.
22. Wang H-x, Zhou Y, Tang C-x, Wu J-g, Chen Y, Jiang Q-w: **Association between bisphenol A exposure and body mass index in Chinese school children: a cross-sectional study**. *Environmental Health* 2012, **11**(1):79.
23. Guyatt GH, Oxman AD, Sultan S, Glasziou P, Akl EA, Alonso-Coello P, Atkins D, Kunz R, Brozek J, Montori V *et al*: **GRADE guidelines: 9. Rating up the quality of evidence**. *J Clin Epidemiol* 2011, **64**(12):1311-1316.
24. Doll R, Hill AB: **Smoking and carcinoma of the lung**. *British medical journal* 1950, **2**(4682):739.

25. Doll R, Hill AB: **Mortality in relation to smoking: ten years' observations of British doctors**. *British medical journal* 1964, **1**(5395):1399.
26. Bross ID: **Spurious effects from an extraneous variable**. *Journal of chronic diseases* 1966, **19**(6):637-647.
27. Blair A, Stewart P, Lubin JH, Forastiere F: **Methodological issues regarding confounding and exposure misclassification in epidemiological studies of occupational exposures**. *American journal of industrial medicine* 2007, **50**(3):199-207.
28. Cox KJ, Porucznik CA, Anderson DJ, Brozek EM, Szczotka KM, Bailey NM, Wilkins DG, Stanford JB: **Exposure classification and temporal variability in urinary bisphenol A concentrations among couples in Utah—the HOPE study**. *Environmental health perspectives* 2016, **124**(4):498.
29. Higgins J, Green S: **Cochrane Handbook for Systematic Reviews of Interventions. Version 5.1.0 (updated March 2011).** **http://handbook.cochrane.org/ [accessed 3 February 2013]**. 2011.

Figures

Figure 1.

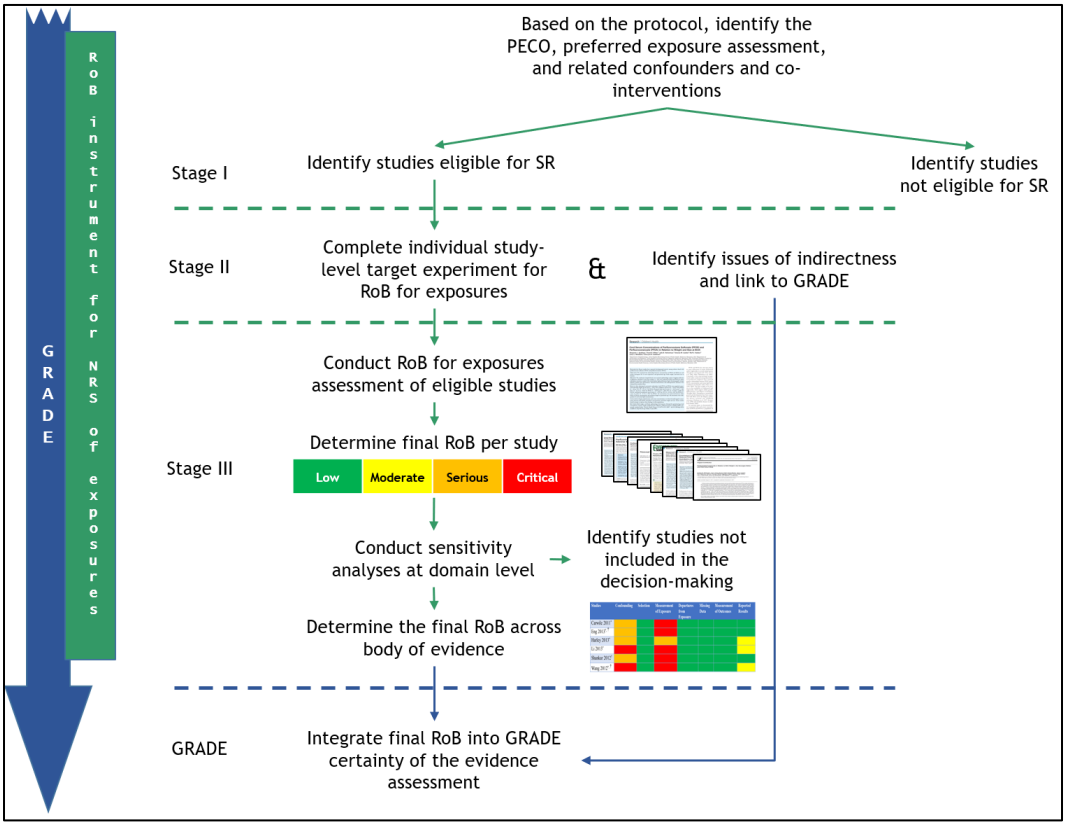


Figure 1. Approach for conducting an assessment using the RoB instrument for NRS of exposures and the integration into GRADE when conducting systematic reviews of exposure.

Tables

Table 1.

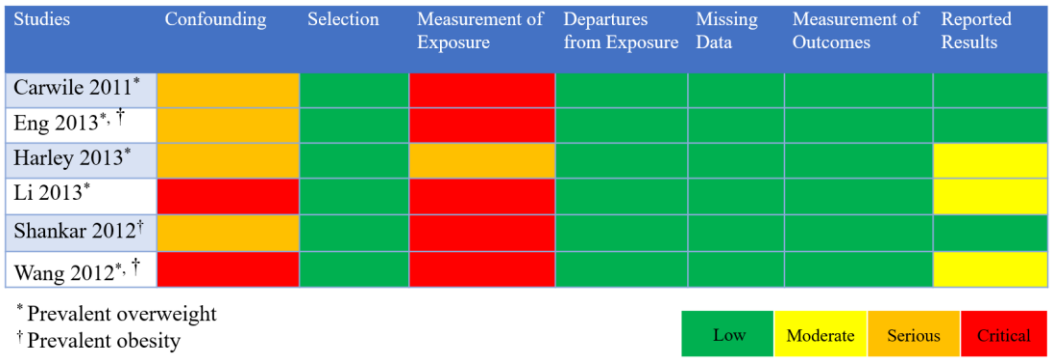


Table 1. Risk of bias matrix presenting judgments for exposure to highest BPA versus exposure to lowest BPA on the outcome of weight.

Table 2.

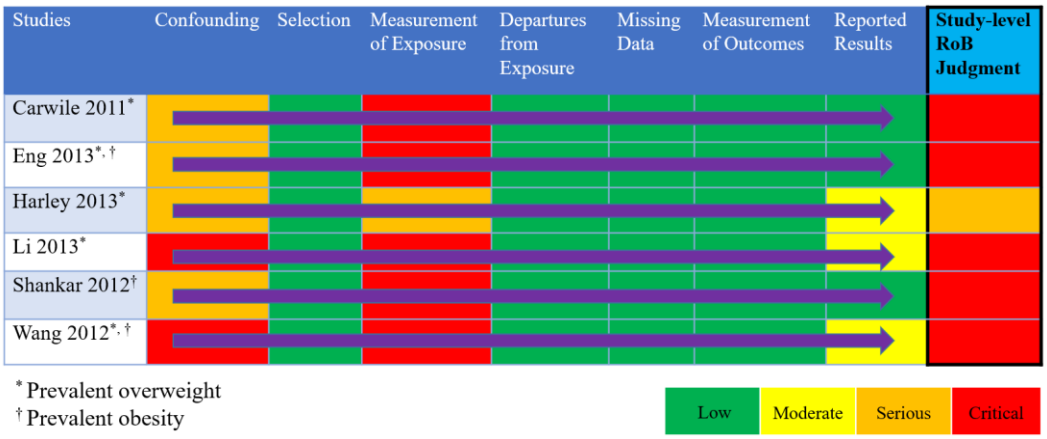


Table 2. Risk of bias matrix presenting study-level judgments for exposure to highest BPA versus exposure to lowest BPA on the outcome of prevalent overweight and prevalent obesity.

Table 3.

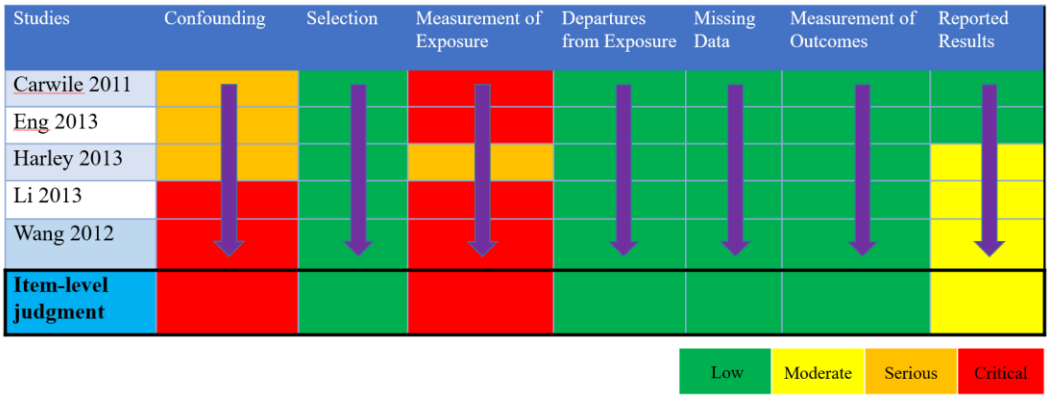


Table 3. Risk of bias matrix presenting item-level judgments for exposure to highest BPA versus exposure to lowest BPA on the outcome of prevalent overweight.

Table 4.

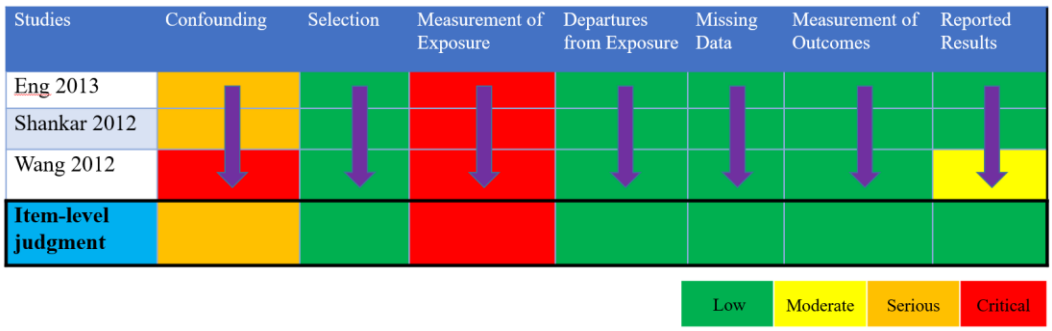


Table 4. Risk of bias matrix presenting item-level judgments for exposure to highest BPA versus exposure to lowest BPA on the outcome of prevalent obesity.

Table 5.

Question: Exposure to highest levels of BPA (CAS# 80-05-7) compared to exposure to lowest levels of BPA in general population
Setting: Community
Bibliography: Rancière, F., Lyons, J. G., Loh, V. H., Botton, J., Galloway, T., Wang, T., ... & Magliano, D. J. (2015). Bisphenol A and the risk of cardiometabolic disorders: a systematic review with meta-analysis of the epidemiological evidence. Environmental Health, 14(1), 46.

Quality assessment							№ of patients		Effect		Quality	Importance
№ of studies	Study design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	exposure to highest BPA levels	exposure to lowest BPA levels	Relative (95% CI)	Absolute (95% CI)		
Prevalent overweight (assessed with: BMI ≥85th percentile for age/gender in children; BMI 18.5-25/30 kg/m2)												
5	studies	very, very serious [a]	not serious [b]	not serious [c]	serious [d]	none	1774/5403 (32.8%)	1584/5657 (28.0%)	**OR 1.21** (0.98 to 1.56)	**40 more per 1,000** (from 4 fewer to 98 more)	⊕◯◯◯ VERY LOW	CRITICAL
Prevalent obesity (assessed with: BMI ≥95th percentile for age/gender in children; BMI ≥25-30 kg/m2)												
3	studies	very serious [a]	not serious	not serious [c]	not serious	none	1425/5178 (27.5%)	1204/5342 (22.5%)	**OR 1.67** (1.32 to 1.93)	**102 more per 1,000** (from 52 more to 134 more)	⊕⊕◯◯ LOW	CRITICAL

CI: Confidence interval; **OR**: Odds ratio

Explanations

a. Most studies adjusted for known confounders of weight (age and gender) and diet; however, two studies did not account for caloric intake or diet which is relevant for evaluating weight-related outcomes, there is some risk of unmeasured confounding; BPA measurement present potential for bias as the chemical is non-persistent with a short half-life and exposure measurements were not repeated (except in one study), one study measures BPA three months post-BMI measurement, remaining studies measure BPA and BMI at the same time; however, the effect estimates may underestimate the true effect reducing our concern of non-differential misclassification; potential risk of reporting bias because three studies did not report prior publication of a protocol; however, all studies present outcome measures and analyses consistent with a priori plan outlined in the manuscript.
b. The I^2 value = 45% and exploration of the forest plot suggests some inconsistency introduced by one outlying study contributing 4.3% of the weight to the analysis of children.
c. Studies measured BPA concentration through urinary output. uBPA (BPA in urine) is considered a reliable and direct measure of BPA consumption and was not downgraded for indirectness.
d. Imprecision is present because the width of the confidence interval is consistent with no association of benefit and a moderate association of benefit and harm.

Table 5. Exposure to BPA on the outcome of birthweight GRADE evidence assessment.

Appendices

Appendix A. Stage I of the RoB instrument for NRS of exposures for the PECO: "What is the effect of highest levels vs. lowest levels of BPA exposure on weight?"

Stage I Items	Response
Confounding for BPA and obesity	• Body composition (age, ethnicity, gender, height, race); • Weight (age, gender); • Waist circumference (age, gender); • Body mass index (age, ethnicity, gender, race); • In addition, consumption of canned or packaged food and drink ("processed" food) that is also energy dense and low-nutrient (e.g., soda) is a significant confounder because food packaging is a main source of exposure to BPA. • Co-exposures: There may be some concern for co-exposure to certain phthalates used in food packaging that have also been linked to obesity. However, phthalates are used in different types of food packaging than BPA (plastic wraps versus canned lining and polycarbonate materials). No other *a priori* co-exposures of particular concern are identified for general population studies. There may be some co-exposures that need to be considered in occupational studies and these should be assessed on a case by case basis if discovered.
Co-interventions	• None identified
Accuracy of the measurement of exposure to BPA (CAS# 80-05-7)	• BPA is a non-persistent compound (near 100% elimination within 24 hours after oral exposure, possible longer elimination time from non-oral exposure but on order of days), so blood and urine measures only assess recent exposure. This means current exposure levels may NOT be indicative of past exposures. This is problematic for assessment of BPA as a risk factor for health outcomes that are not acute and take time to develop like obesity. • BPA measures are variable over time in the same person (even during the same day) so methods that utilize repeated measures of exposure are preferred. Some experts on BPA exposure assessment express less concern for lack of repeated measures for NHANES data because it is a large sample survey of the general population. • Standard analytical measures: Measurement of urine or blood by quantitative techniques such as liquid chromatography-triple quadrupole mass spectrometry (LC-MS/MS) and high-pressure liquid chromatography with tandem mass spectrometry (HPLC/MS) are preferred. Measurements made at CDC are considered high-quality. • Measures to minimize sample contamination with BPA should be taken (e.g., glass pipettes, polypropylene plastic lab ware and sample collection materials, water blanks). • Measures of unconjugated BPA in blood need to be very carefully considered based on extent to which investigators controlled for background exposures.

	• Questionnaire or self-reported measures of BPA exposure are more problematic due to the ubiquity of exposure and lack of knowledge on all possible routes of exposure, e.g., thermal paper, certain pharmaceuticals. However, there is some support for an association between higher urine/blood levels of BPA and higher reported use of BPA-containing food packaging (e.g., canned food consumption) or handling of BPA-containing thermal paper (cashiers) so questionnaire data that assess these types of exposure sources may have some utility in assessing longer-term time trends in exposure.
Accuracy of the measurement of outcome of obesity	• Body Composition: Dual-energy X-Ray absorptiometry, triceps skinfold thickness, subscapular skinfold thickness, suprailiac skinfold thickness • Measured waist circumference • Body mass index • Measured weight *Obesity typically develops relatively slowly over time so preferred follow-up times after start of exposure would be on the order of several months to years.

Appendix B. Stage II of the RoB instrument for NRS of exposures for Carwile & Michels, 2011.

Specify a target randomized trial specific to the study

Design	Individual randomized controlled trial
Participants	Adults of all ages, predominantly 18-35 years (8.2% <18 years and 7.9% > 35 years). Civilian, non-institutionalized, United States population. Analyses restricted to participants 18–74 years of age, who were included in the random subsample of participants, who supplied a spot urine sample analyzed for BPA.
Experimental intervention	BPA highest levels (quartile 4: ≥ 4.7 ng/mL)
Comparator	BPA lowest levels (quartile 1: ≤ 1.1 ng/mL)

Specify the outcome

Specify which outcome is being assessed for risk of bias (typically from among those earmarked for the Summary of Findings table). Specify whether this is a proposed benefit or harm of intervention.

Prevalent overweight (Overweight: 25 ≤ BMI < 30 kg/m^2 [reference: BMI <25 kg/m^2])

Specify the numerical result being assessed

In case of multiple alternative analyses being presented, specify the numeric result (e.g. RR = 1.52 (95% CI 0.83 to 2.77) and/or a reference (e.g., to a table, figure or paragraph) that uniquely defines the result being assessed.

Participants in the upper BPA quartile 4 vs. participants in the lowest BPA quartile 1: OR: 1.76, 95% CI: 1.06–2.94)

(i) Confounding domains listed in Stage I				
Confounding domain	Measured variable(s)	Is there evidence that controlling for this variable was unnecessary?	Is the confounding domain measured validly and reliably by this variable (or these variables)?	OPTIONAL: Is failure to adjust for this variable (alone) expected to favor the experimental intervention or the comparator?
			Yes / No / No information	Favor experimental / Favor comparator / No information
Age, gender	Weight	No	Yes	Favor experimental
Consumption of canned or packaged food and drink ("processed" food) that is also energy dense and low-nutrient (e.g., soda)	Daily caloric intake	No	No	Favor experimental because obese individuals (potentially caused by higher consumption of canned foods and drinks) have higher urinary BPA levels relative to those with normal weight.

(ii) Additional confounding domains relevant to the setting of this particular study, or which the study authors identified as important				
Confounding domain	Measured variable(s)	Is there evidence that controlling for this variable was unnecessary?	Is the confounding domain measured validly and reliably by this variable (or these variables)?	OPTIONAL: Is failure to adjust for this variable (alone) expected to favor the experimental intervention or the comparator?
			Yes / No / No information	Favor experimental / Favor comparator / No information
Alcohol drinking, fish intake, protein, fat, carbohydrate, and energy intake	none	no	no	

Carwile JL, Michels KB: **Urinary bisphenol A and obesity: NHANES 2003–2006**. *Environmental research* 2011, **111**(6):825-830.

Appendix C. Stage II of the RoB instrument for NRS of exposures for Harley et al., 2013.

Specify a target randomized trial specific to the study

Design	Individual randomized controlled trial
Participants	Children at 5 and 9 years of age born to eligible pregnant women were at least 18 years of age, spoke English or Spanish, qualified for low-income health insurance, were at < 20 weeks gestation, and were planning to deliver at the county hospital. Must have had a singleton, live birth.
Experimental intervention	BPA highest levels (tertile 3: 4.6–349.8 μg/g)
Comparator	BPA lowest levels (tertile 1: <LOD-2.4 μg/g)

Specify the outcome

Specify which outcome is being assessed for risk of bias (typically from among those earmarked for the Summary of Findings table). Specify whether this is a proposed benefit or harm of intervention.

Prevalent overweight (Overweight: BMI ≥ 85th percentile at 5 and 9 years of age)

Specify the numerical result being assessed

In case of multiple alternative analyses being presented, specify the numeric result (e.g. RR = 1.52 (95% CI 0.83 to 2.77) and/or a reference (e.g. to a table, figure or paragraph) that uniquely defines the result being assessed.

Participants in the upper BPA tertile 3 vs. participants in the lowest BPA tertile 1: OR = 1.36 (0.75–2.47)

(i) Confounding domains listed in Stage I				
Confounding domain	Measured variable(s)	Is there evidence that controlling for this variable was unnecessary?	Is the confounding domain measured validly and reliably by this variable (or these variables)?	OPTIONAL: Is failure to adjust for this variable (alone) expected to favor the experimental intervention or the comparator?
			Yes / No / No information	Favor experimental / Favor comparator / No information
Age, gender	Weight	No	Yes	Favor experimental
Consumption of canned or packaged food and drink ("processed" food) that is also energy dense and low-nutrient (e.g., soda)	Child consumption of soda, fast food, and sweets	No	Yes	Favor experimental because obese individuals (potentially caused by higher consumption of canned foods and drinks) have higher urinary BPA levels relative to those with normal weight.

(ii) Additional confounding domains relevant to the setting of this particular study, or which the study authors identified as important				
Confounding domain	Measured variable(s)	Is there evidence that controlling for this variable was unnecessary?	Is the confounding domain measured validly and reliably by this variable (or these variables)?	OPTIONAL: Is failure to adjust for this variable (alone) expected to favor the experimental intervention or the comparator?
			Yes / No / No information	Favor experimental / Favor comparator / No information
Television watching	Average daily TV time	No	Yes	Favor experimental
Environmental tobacco smoke exposure	Self-reported mother's smoking status	No	Yes	No information
Time spent playing outdoors	Unknown	No	No information	No information

Harley KG, Schall RA, Chevrier J, Tyler K, Aguirre H, Bradman A, Holland NT, Lustig RH, Calafat AM, Eskenazi B: **Prenatal and postnatal bisphenol A exposure and body mass index in childhood in the CHAMACOS cohort**. *Environmental health perspectives* 2013, **121**(4):514.

Appendix D. Summary of Stage III of the RoB instrument for NRS of exposures and the direction of bias and reaching the overall bias judgement for Carwile & Michels, 2013.

Bias items	Risk of bias	Direction of bias	Rationale
Bias due to confounding	Serious	Unknown	NHANES data were used. Specific details were not provided in the study report, but NHANES co-variate data were obtained from either a standardized questionnaire or laboratory methods (e.g., creatinine). The reliability/validity of the questionnaire was not reported, but it is not expected to appreciably bias the results. Most of the critical confounders were considered statistically, but there is possibility of residual unmeasured (and unidentified) confounding. For the most part, although certain post-exposure variables are relevant to evaluating obesity (e.g., caloric intake), there is little information on the association of these variables to BPA exposure. No indication that time-varying confounding is a major concern given the cross-sectional nature of the study. Critical confounders (age, gender, and ethnicity) were accounted for in the analysis. Model 1 was adjusted for age, sex, and urinary creatinine. Model 2 was adjusted for race, education, and smoking in addition to Model 1 covariates.
Bias in selection of participants into the study	Low	N/A	Study is cross-sectional. Subjects were randomly selected from NHANES subjects with urinary BPA data available using the same criteria. Selection of subjects was unrelated to either exposure or outcome. While there is no information on start of exposure, everyone is exposed to BPA throughout their life, but the levels will change over time. Although BPA is ubiquitous, start of exposure and how exposure changes over time are not known. Timing of recruitment was similar (2003-2006), but given that the age ranged from 18 to 74 years, exposure could range by more than a decade.
Bias in classification of exposures	Critical	Concerns of bias toward the null due to non-differential misclassifica	Urinary BPA concentration was measured in 1 spot sample from each participant. The lower limit of detection (LLOD) was 0.36ng/ml in 2003/04 and 0.4ng/ml in2005/06. For BPA concentrations below the LLOD (2003/04: n=110/1373 [8%]; 2005/06: n=114/1374 [8%]) NHANES assigned a value of the LLOD divided by the square root of two. BPA is a

	tion of the exposure.	non-persistent compound and exposure measures were not repeated. Therefore, there is no confidence that the current exposure reflects exposure over the subject's life time or even over any duration of time. Because this population is obtained from NHANES some experts consider the lack of repeated measures to be less of a concern because it is a large survey of the general population (this cross-sectional study had a population of 2747 adults). Exposure was measured at same time as outcome, but participants were likely exposed throughout life due to BPA being a ubiquitous exposure. Therefore, it is unlikely that entry into the cohort started with the exposure. Cross-sectional analyses with both BPA exposure and weight, height, and waist circumference used to define obesity assessed simultaneously. Urine samples were obtained at the time that obesity measurements were obtained and analyzed later in a laboratory separate from where the data were collected. In addition, NHANES collected data on a variety of compounds and health effects without knowledge of the intent for this current study indicating that exposure status is not likely to be biased by knowledge of the outcome. The range/variability in exposure was likely sufficient with a 25th to 75th percentile range of 1.18 to 3.33 ng/mL urinary BPA ng/mL and quartiles ranging from <1.1 ng/mL to >4.7 ng/mL. However, we are not confident that the subjects were exposed to this concentration for a long period of time. Lacking information on the duration that subjects were exposed to these levels, the single BPA measurement obtained at the same time as outcome is not of sufficient to detect an effect of exposure. Urinary BPA samples were collected at the same time that height, weight, and waist circumference were measured. Because BPA is not persistent and obesity is not an acute effect, there is not adequate follow-up period to allow for the development of the outcome of interest. Total (free and conjugated) urinary BPA concentrations were measured at the Division of Environmental Health Laboratory Sciences (National Center for Environmental Health, CDC) using online solid-phase extraction coupled to

			isotope dilution high-performance liquid chromatography–tandem mass spectrometry. Quality control (QC) procedures included analysis of reagent blanks and samples of pooled human urine spiked with BPA at low-and high-concentrations. Coefficients of variation calculated for low-and high-concentration QC samples were 19% and 12% in 2003–2004 and 13% and 11% in 2005–2006. Additional information on laboratory methods is available online (CDC, 2004b, 2006b).
Bias due to deviations from intended exposures	Low	N/A	There is little concern that changes in exposure status occurred among participants. Although BPA levels may change overtime, the cross-sectional nature of the study and the intention-to-treat analyses this is of little concern because participants are analyzed based on the exposure group they are assigned from the single measurement. No critical co-exposures were identified and nothing about the subject characteristics suggests likelihood of differential exposure to other environmental contaminants at lower versus higher concentrations of BPA.
Bias due to missing data	Low	N/A	There is no information on the missing data by exposure level, but it is unlikely to be related to exposure level. The missing indicator method was used for covariates with missing data for >=10% of observations, otherwise observations with missing covariate data were excluded. Data excluded from analysis did not exceed 4% and is considered relatively complete. 32 or 87 observations were stated excluded from analysis due to missing BMI data depending on the analysis conducted. 47 participants were excluded based on missing urinary BPA measurements. There were observations excluded based on missing covariate data. The number varied with the analysis, but was only excluded if it was <10%.
Bias in measurement of the outcome	Low	N/A	It is unlikely that the outcome could be affected by knowledge of exposure. Height, weight, and waist circumference were measured using standard NHANES protocols (not described in the publication, but available on NHANES website). Body mass index was calculated (weight (kg)/height $(m)^2$). The specific measurements would not be affected by knowledge of exposure, and it is unlikely that the calculation or assignment into obesity category would be affected by knowledge of exposure.

			Specific methods were not reported in the study report, but are provided on NHANES website. Height and weight are likely sensitive measurements with waist circumference likely slightly less sensitive. Height, weight, and waist circumference were measured by trained technicians using a standardized protocol. Method details, including QA/QC procedures, are available on the NHANES website. BMI was calculated as weight in kilograms divided by height in meters squared and used to define overweight [25.0 <BMI<29.9] and obesity [BMI >30.0]. It is unlikely that any systematic error in measuring height, weight, or waist circumference (or in calculating the BMI or assigning obesity category) would have been related to exposure. NHANES has a standard protocol for measuring height, weight, and waist circumference that would have been used for all subjects. Outcome was assessed at the time of sample collection for exposure. Therefore, exposure was unknown at time of outcome assessment.
Bias in selection of the reported result	Low	N/A	Reporting of the results is consistent with an a priori plan and data were readily available from NHANES that provides all protocols for obtaining the data online. Results were provided for two measurements of obesity, which were reported in the methods making it unlikely that there is selective reporting based on outcome. Statistical methods reported in the methods section were used and presented in the results. Associations between urinary BPA and obesity were assessed for effect modification by gender, which were provided in the supplemental material.
Overall bias	**Serious**	**Possibly toward the null**	Overall bias was judged as Serious due to concerns of potential unknown confounders, unmeasured confounding due to the single time-point data collection, and concerns of non-differential misclassification of the exposure.

Carwile JL, Michels KB: **Urinary bisphenol A and obesity: NHANES 2003–2006**. *Environmental research* 2011, **111**(6):825-830.

Appendix E. Summary of Stage III of the RoB instrument for NRS of exposures and the direction of bias and reaching the overall bias judgement for Harley et al., 2013.

Bias items	Risk of bias	Direction of bias	Rationale
Bias due to confounding	Moderate	Unknown	Most of the critical confounders were considered statistically, but there is possibility of residual unmeasured (e.g., diet, pesticide exposure) confounding. The study evaluated the child's BPA exposure throughout several points in their life. And used each one separately in the evaluation. Changes in BPA exposure could be related to changes in food consumption over time as BPA exposure is mainly through canned or processed food including soda, which could also be related to obesity. Since Harley follows participants over time, there is some concern for time-varying confounding as they may have changed their diet while pregnant. Potential confounders were identified a priori using directed acyclic graphs. Potential confounders included maternal pre-pregnancy BMI, age, education, years of residence in the United States, smoking during pregnancy, soda consumption during pregnancy, and family income. Time-varying covariates considered were child consumption of soda, fast food, and sweets, television watching, environmental tobacco smoke exposure, and time spent playing outdoors, assessed at multiple times during childhood. Covariates were included in the final models if they were associated with both exposure and any of the growth outcomes at p-value < 0.2 or if removing them changed the coefficient for the main BPA exposure variable by > 10%. Maternal age and pre-pregnancy BMI were analyzed as continuous variables. Other variables were categorical. Mothers were interviewed twice during pregnancy, after delivery, and when their children were 2, 3.5, 5, 7, and 9 years of age to obtain information about demographic characteristics, diet, and behaviors. All interviews were conducted in English or Spanish using structured questionnaires, but no information was provided on reliability/validity.

			At the baseline interview, we asked mothers about their race/ethnicity, education, income, marital status, and number of years they had lived in the United States, as well as information about soda consumption, smoking, and alcohol and drug use during pregnancy. We calculated pre-pregnancy BMI from self-reported pre-pregnancy weight and measured height. If self-reported pre-pregnancy weight was unavailable or invalid, we used measured weight at first prenatal visit (n = 23) if the first prenatal visit occurred at or before 13 weeks gestation or used regression models to impute pre-pregnancy weight based on weight at all prenatal visits if the first prenatal visit occurred after 13 weeks (n = 16).
Bias in selection of participants into the study	Low	N/A	Selection of subjects was unrelated to either exposure or outcome. The study sample consisted of participants in the Center for the Health Assessment of Mothers and Children of Salinas (CHAMACOS), a longitudinal cohort study of environmental factors and children's growth and development. Pregnant mothers were enrolled Selection of subjects was unrelated to either exposure or outcome in 1999 and 2000 from prenatal clinics serving the farmworker population in the Salinas Valley, California. Eligible women were at least 18 years of age, spoke English or Spanish, qualified for low-income health insurance, were at < 20 weeks gestation, and were planning to deliver at the county hospital. Mothers provided written informed consent for themselves and their children to participate in the study. Start of exposure occurred in the first trimester and all subjects were followed through 9 years of age.
Bias in classification of exposures	Serious	Some concern of bias toward the null due to non-differential misclassificati on of the exposure.	Urinary BPA concentration was measured in 4 spot samples, 2 during pregnancy and 2 from the child. LOD was 0.4 ng/mL. Concentrations < LOD for which a signal was detected were reported as measured. Concentrations < LOD with no signal detected were randomly imputed based on a log-normal probability distribution using maximum likelihood estimation. Initial exposure was measured during the first trimester of pregnancy. While this may not be the exact date of start of exposure it would be pretty close for the children.

	Prenatal and five-year-old exposure measurements were taken prior to the assessment of BMI at 9 years. Exposure was assessed prior to the outcome at three different time points. Only one exposure measurement was obtained at the same time as the outcome. Thus, it was not possible for classification of exposure to have been affected by the knowledge of the outcome. The range/variability in exposure was sufficient (range during pregnancy 0.5 to 4.6 ng/mL and during childhood 0.9 to 16.3 ng/mL). Although BPA levels change over time and we are not confident that the subjects were exposed to this concentration for a long period of time, the fact that there were 4 measurements per subject make us more confident in the exposure being represented of changes over time. In addition, since the child's exposure was first measured based on mother's levels when pregnant, then again when the children were 5 (4 years prior to measuring outcome) the duration of exposure would have been sufficient even if the level of this exposure was not consistent. BPA levels were also measured in the child at 9 years. However, data were not provided for the individual subjects to know how the BPA levels may have varied per subject. Children were followed up for 9 years, which would have been sufficient time for the outcome to develop. Spot urine samples were collected from mothers at two timepoints during pregnancy: near the end of the first (mean ± SD, 13.8 ± 5.0 weeks gestation) and second (mean ± SD, 26.4 ± 2.4 weeks gestation) trimester of pregnancy and from the children when they were 5 (mean ± SD, 5.1 ± 0.2 years) and 9 (mean ± SD, 9.4 ± 0.4 years) years of age. Urine samples were collected in polypropylene urine cups, aliquoted into glass vials, and frozen at –80°C until shipment to the CDC for analysis. Analysis of field blanks showed no detectable contamination by BPA using this collection protocol. Solid-phase extraction coupled to high performance liquid chromatography–isotope dilution tandem mass spectrometry to measure total urinary BPA concentration (conjugated

			plus unconjugated). Concentrations < LOD for which a signal was detected were reported as measured. Concentrations < LOD with no signal detected were randomly imputed based on a log-normal probability distribution using maximum likelihood estimation. Specific gravity was measured with a refractometer (National Instrument Company Inc., Baltimore, MD) for the maternal urine samples, but was unavailable for the children's samples. Thus, maternal concentrations were normalized for urinary dilution using urine specific gravity, and child BPA concentrations were normalized by dividing by urinary creatinine concentration.
Bias due to deviations from intended exposures	Low	N/A	There is little concern that changes in exposure status occurred among participants. Although BPA levels may change overtime, several measurements were obtained and evaluate separately by exposure they were assigned. Because each exposure was evaluated as an intent to treat, there is little concern about the potential changes in exposure. The study authors reanalyzed the models controlling separately for three important prenatal exposures in this population: organochlorine pesticides [using prenatal serum concentrations of dichlorodiphenyldichloroethylene (DDE)], organophosphate pesticides (using prenatal urinary metabolites of organophosphate pesticides), and brominated flame retardants [using prenatal serum concentrations of polybrominated diphenyl ethers (PBDEs)].
Bias due to missing data	Low	N/A	Reasons for exclusion were documented and unlikely to differ across exposures threshold. Although some subjects were lost to follow-up and the missing data were not described by exposure status, the study authors conducted analyses that addressed loss to follow-up and are likely to have removed any risk of bias thus judged low risk of bias. There is no statement that participants with missing covariate data were excluded from analyses. There is no information on the missing data by exposure level. Although it is unlikely to be related to exposure level, they had the data in order to compare those lost to follow-up with those included in the analysis, but no information was provided.

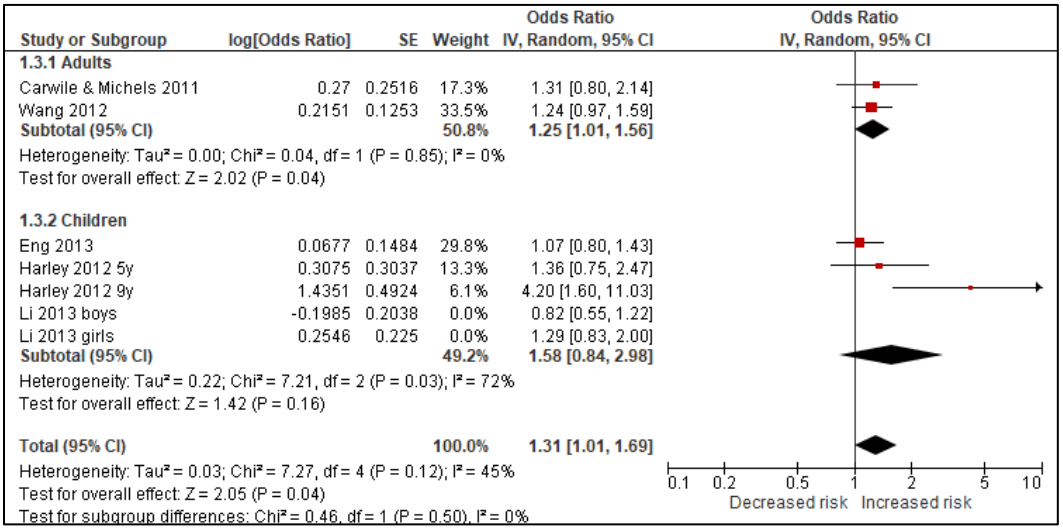
			Of the 527 mothers meeting the inclusion criteria, 402 had at least one urine measurement available. There were 325 measurements in children at 5 years and 304 available at 9 years. Of the 402 children included in the analysis, anthropometric measurements were available for 319 children at 5 years and 311 children at 9 years.
Bias in measurement of the outcome	Low	N/A	It is unlikely that the outcome could be affected by knowledge of exposure. It was not noted that outcome assessors were blind to the exposure level, but it was likely given that separate individuals were used to measure the outcome parameters than conducted the exposure analysis (i.e., CDC). The same methods were used for all participants at all times measured. It is unlikely that any systematic error in anthropometric measurements (or calculating the BMI or assigning obesity category) would have been related to exposure. Children were weighed and measured without jackets or shoes by trained study staff. Weight was measured using a digital scale and rounded to the nearest 0.1 kg. Height was measured using a stadiometer and rounded to the nearest 0.1 cm. Starting at 5 years of age, waist circumference was measured at each visit by placing a measuring tape around the abdomen at the level of the iliac crest, parallel to the floor. Height and waist circumference measurements were conducted in triplicate and averaged for analysis. When the children were 9 years of age, fat percentage was measured using "foot-to-foot" bio-impedance technology with a Tanita TBF-300A body composition analyzer (Tanita Corp.). BMI was calculated as weight (kilograms) divided by height squared (square meters) and compared with the sex-specific BMI-for-age percentile data issued by CDC in 2000 (National Center for Health Statistics 2005). Children who were ≥ 85th but < 95th percentile for their age and sex were classified as overweight. Age- and sex-standardized BMI z-scores were also generated using the CDC norms. These methods are considered sensitive.
Bias in selection of	Moderate	Potential for bias away from the null.	Reported results are consistent with an a priori plan; however, as no protocol was published prior to the study there is potential for

the reported result			reporting bias to inflate results for publication success. Several measurements of obesity were evaluated and reported. These were also assessed at several different time periods in the children. Although the publication only shows a few of the results (both positive and negative), the BMI-z-scores for all ages are presented in the supplemental data indicating that it is unlikely that there was bias from selective reporting of outcome. Gender and age were evaluated as separate subgroups as described in the report. Statistical methods reported in the methods section were used and presented in the results or discussion. BPA was analyzed as categorical and continuous variable.
Overall bias	**Moderate**	**Unknown**	Overall bias was judged as Moderate due to concerns of potential unknown confounders, some concerns of non-differential misclassification of the exposure, and some concerns with bias in reported results.

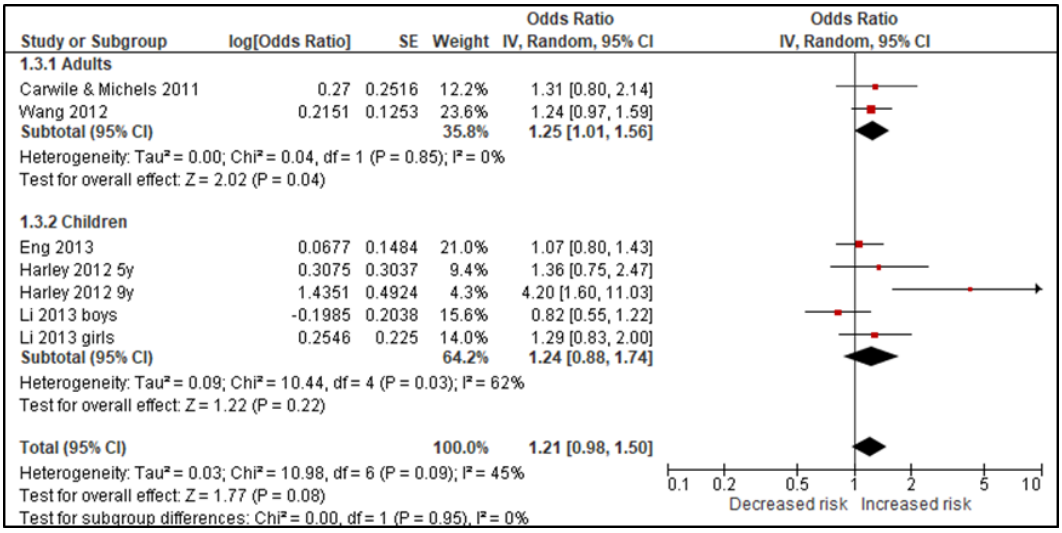
Harley KG, Schall RA, Chevrier J, Tyler K, Aguirre H, Bradman A, Holland NT, Lustig RH, Calafat AM, Eskenazi B: **Prenatal and postnatal bisphenol A exposure and body mass index in childhood in the CHAMACOS cohort**. *Environmental health perspectives* 2013, **121**(4):514.

Appendix F. Sensitivity analysis for the outcome of prevalent overweight.

Figure F.1. Sensitivity analysis of studies with 'Serious' bias due to confounding

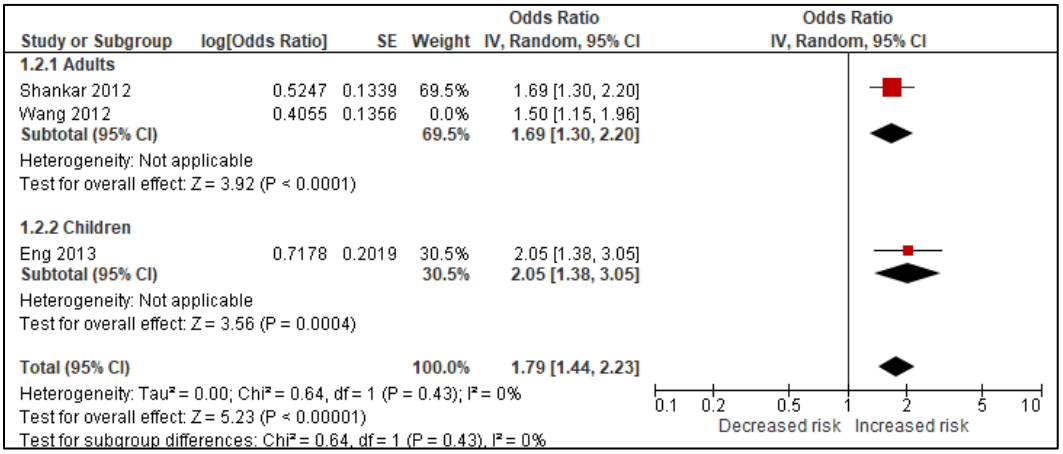


F.2. Sensitivity analysis of all studies

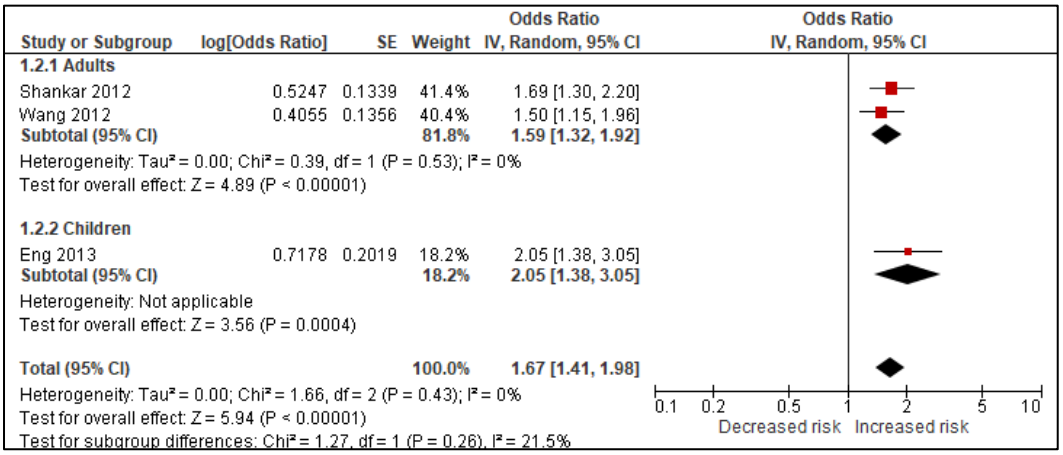


Appendix G. Sensitivity analysis for the outcome of prevalent obesity.

G.1. Sensitivity analysis of studies with 'Serious' bias due to confounding



G.2. Sensitivity analysis of all studies



CHAPTER 5. RELIABILITY AND VALIDITY OF RISK OF BIAS INSTRUMENTS IN STUDIES OF ENVIRONMENTAL EXPOSURES

PREFACE TO CHAPTER 5

Chapter 5. *Reliability and validity of risk of bias instruments in studies of environmental exposures* has been reviewed by all co-authors and will be submitted to Environment International.

Reliability and validity of risk of bias instruments in studies of environmental exposures

Author list

Rebecca L. Morgan [a]; Kristina A. Thayer [b]; Sandra Montiero [a]; Nancy Santesso [a]; Alison C. Holloway [c]; Robyn Blain [d]; Alexandra E. Goldstone [d]; Pam Ross [d]; Holger J. Schünemann [a,e]

Affiliations

[a] Department of Health Research Methods, Evidence, and Impact (formerly the Department of Clinical Epidemiology & Biostatistics), McMaster University, Health Sciences Centre, Room 2C14, 1280 Main Street West, Hamilton, ON L8S 4K1 Canada morganrl@mcmaster.ca, monteisd@mcmaster.ca, santesna@mcmaster.ca, schuneh@mcmaster.ca

[b] Division of the National Toxicology Program, National Institute of Environmental Health Sciences, National Institutes of Health, Department of Health and Human Services, P.O. Box 12233, Mail Drop K2-02, Research Triangle Park, NC 27709 USA. thayer@niehs.nih.gov

[c] Department of Obstetrics and Gynecology, McMaster University, Health Sciences Centre, Room 3N52A, 1280 Main Street West, Hamilton, ON L8S 4K1 Canada. hollow@mcmaster.ca

[d] ICF, 9300 Lee Highway, Fairfax, VA 22031 USA. Robyn.Blain@icf.com, Pam.Ross@icf.com, Ali.Goldstone@icf.com

[e] Department of Medicine, McMaster University, Health Sciences Centre, Room 2C14, 1280 Main Street West, Hamilton, ON L8S 4K1 Canada. schuneh@mcmaster.ca

Corresponding author: Holger J. Schünemann. Department of Health Research Methods, Evidence, and Impact (formerly the Department of Clinical Epidemiology & Biostatistics), Health Sciences Centre, Room 2C14, 1280 Main Street West, Hamilton, ON L8S 4K1 Canada. schuneh@mcmaster.ca.

Conflict of interest

The authors declare they have no competing financial interests with respect to this manuscript, or its content, or subject matter.

Abstract

Background: The Risk of bias (RoB) instrument for non-randomized studies (NRS) of exposures is a new tool that evaluates RoB of NRS on seven domains using a standardized comparison to a randomized target experiment. This instrument provides a more detailed RoB assessment than existing instruments.

Objectives: To assess reliability and validity of the RoB instrument for NRS of exposures through comparison with other RoB instruments for exposure studies and topic experts' judgments.

Methods: We evaluated interrater reliability of the RoB instrument for NRS of exposures in 35 studies with three raters and three commonly-used RoB instruments for exposure studies: Newcastle-Ottawa Scale (NOS), and tools used by the National Toxicology Programs' Office of Health Assessment and Translation (OHAT), and Office of the Report of Carcinogens (ORoC). We also assessed the instruments' validity by calculating correlation coefficients between instruments and burden comparing it against 28 experts' global judgment of RoB.

Results: The RoB instrument for NRS of exposures demonstrates substantial interrater reliability (intraclass correlations, ICC = 0.73; 95% CI: 0.53-0.85). Other instruments showed comparable ICCs: NOS = 0.89 (95% CI: 0.80-0.94); OHAT = 0.80 (95% CI: 0.64-0.89); and ORoC = 0.70 (95% CI: 0.47-0.84). The RoB instrument for NRS of exposures also appears valid (r = 0.71 to 0.90) compared to ratings on other RoB instruments and

28 topic expert global judgements. However, the RoB instrument for NRS of exposures requires more time for completion.

Conclusions: The RoB instrument for NRS of exposures is reliable and valid and may allow for more informed and detailed RoB judgments than other instruments.

Keywords (6): Risk of bias, ROBINS, GRADE, non-randomized study, reliability, validity

Highlights

- Understanding an instrument's measurement properties should influence systematic review authors' decisions when choosing a RoB instrument to use for study evaluation.
- The RoB instrument for NRS of exposures instrument includes distinct and relevant RoB constructs to which exposure studies are prone and provides detailed signalling questions to facilitate the assessment.
- The RoB instrument for NRS of exposures instrument demonstrates substantial interrater and inter-instrument reliability and is ready for wider adoption by systematic review authors.

Abbreviations

AEC: absolute error coefficient

ANOVA: analysis of variance

BPA: bisphenol-A

CI: confidence interval

EPA: U.S. Environmental Protection Agency

G theory: generalizability theory

GRADE: Grading of Recommendations Assessment, Development, and Evaluation

ICC: intraclass correlation

IRIS: Integrated Risk Information System

NOS: Newcastle-Ottawa Scale

NRS: non-randomized studies

NTP: National Toxicology Program

OHAT: Office of Health Assessment and Technology

ORoC: Office on the Report on Carcinogens

PBDE: polybrominated diphenyl ethers

PFOA: perfluorooctanoic acid

$PM_{2.5}$: particulate matter with aerodynamic diameter less than 2.5 μm

RCT: randomized controlled trial

RoB: risk of bias

ROBINS-I: Risk of Bias in Non-randomized Studies of Interventions

TSH: thyroid simulation hormones

1. Introduction

Multiple instruments are in use to assess the risk of bias (RoB), often called study quality or internal validity, in non-randomized studies (NRS) [1, 2]. RoB reflects the potential of limitations in a study to cause a systematic deviation of the results (either inflated or underestimated) from the true effect. Due to a lack of clear advantages of one instrument over another, no single instrument is recommended strongly for use in systematic review or guideline development [3, 4]. The lack of guidance on what instrument to use is a key issue in environmental health, where NRS of exposures predominate. There also is a need to collect empirical data to establish the importance of individual RoB domains and the measurement properties of the instruments [5].

There also is a desire by systematic review authors in the environmental health field to change the method of integration of NRS into knowledge synthesis and decision-making, specifically in the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) approach [6]. Currently, NRS are rated down to an initial rating of low certainty because of the potential for confounding and selection bias causing an imbalance of prognostic characteristics in the study population and the exposures of interest which randomization attempts to prevent. The GRADE Working Group recognizes that a RoB instrument evaluating NRS in a standardized comparison against randomized controlled trials (RCTs) would allow all bodies of evidence to begin with a high certainty rating [7]. In environmental health, this comparison against an RCT may be expressed as a standardized comparison to a randomized target experiment [8].

We developed the RoB instrument for NRS of exposures in response to the feedback from systematic review authors for a more detailed assessment that integrates with the GRADE approach [8]. The instrument resulted from multiple pilot tests and external feedback that suggested adaptations to the Risk of Bias Instrument for Non-randomized Studies for Interventions (ROBINS-I). ROBINS-I evaluates RoB in studies of health interventions using a standardized comparison to RCTs [9]. The RoB instrument for NRS of exposures asks raters to compare studies to a (hypothetical) ideal randomized target experiment. Responses to signaling questions help assess the potential for bias across seven domains: 1) bias due to confounding, 2) bias in selection of participants into the study, 3) bias in classification of exposures, 4) bias due to departures from intended interventions, 5) bias due to missing data, 6) bias in measurement of outcomes, and 7) bias in selection of reported results. The signaling questions are very detailed and break down the specific concepts to inform and aid in the assessment. For example, within the domain bias due to missing data, instead of simply asking the assessor if the presence of too much missing data would lead to potential bias, assessors are asked a series of questions: 1) “Were there missing data?”; 2) “Were participants excluded due to missing data on exposure status?”; 3) “Were participants excluded due to missing data on other variables needed for analysis?”. Then, if these situations are present, the rater decides “Are the proportions of participants and reasons for missing data similar across exposures?” and “Were appropriate statistical methods used to account for missing data?” Only after considering all of the above can raters make a fully-informed decision about whether or not there is bias due to missing data.

For each study, the degree of bias is based on the most conservative (worst) of the domain-level RoB judgments. Study-level bias is identified as 'Low,' 'Moderate,' 'Serious,' or 'Critical.' We developed the RoB instrument for NRS of exposures with topic-specific experts in environmental exposures and epidemiologists so that it accurately measured bias common to studies of exposures [8]. Since there is no reference standard RoB instrument for exposure studies (i.e., gold standard), these suggestions and feedback served to establish a degree of face validity [10].

A number of other RoB instruments are being used in systematic reviews conducted in environmental health, including The National Toxicology Program's (NTP) Office of Health Assessment and Translation (OHAT); the University of California in San Francisco's Navigation Guide; the NTP's Office of the Report on Carcinogens (ORoC); and the Integrated Risk Information System of the U.S. Environmental Protection Agency (EPA-IRIS) [8, 11-14]. The OHAT and Navigation Guide instruments use study-design driven approaches and signaling questions for assessing RoB in individual human and animal studies and are essentially the same instrument [12, 14]. The RoB instrument used by the ORoC to evaluate human evidence is similar to EPA-IRIS and includes a 'study sensitivity' domain [11, 13, 15]. The Newcastle-Ottawa Scale (NOS) is an instrument designed for the purposes of RoB assessment in NRS of health interventions. While not designed with the explicit purpose of evaluating studies of exposure, NOS is frequently used to evaluate cohort, case-control, and cross-sectional studies of exposure including air quality, occupational exposure, and toxins [16-18].

None of the instruments use a comparison against a randomized experiment as the possible reference for a (hypothetical) least biased study. However, many similarities exist between these instruments at the domain and question level, such as tailoring of the instrument to the design of the study (e.g., non-randomized or experimental animal), and using the GRADE approach to assess RoB on an outcome level [5]. Also, the ideas in the domains of assessment (e.g., participant selection, confounding, attrition/exclusion, exposure/intervention assessment, outcome assessment and selective reporting) merge. However, a few differences exist between the instruments, namely in the specific items used for RoB assessment, whether or not to provide a RoB judgment across a body of evidence, and the procedures for evaluating RoB across studies in a systematic review. In addition, we are not aware of published evidence of reliability or validity testing for any instrument designed for the purposes of RoB assessment of exposure studies.

In this study, we assessed the reliability and validity of the RoB instrument for NRS of exposures and compared it with the OHAT/Navigation Guide, 2) ORoC/EPA-IRIS, and 3) NOS. Specifically, we first evaluated the interrater and inter-instrument reliability of each instrument by assessing individual studies across seven case-study topics. Second, we evaluated construct validity of the RoB instrument for NRS of exposures and compared those measurement properties with results from the OHAT, ORoC, and NOS instruments. Results from the reliability and validity analyses can inform the wider adoption of the RoB instrument for NRS of exposures and support potential improvements of bias measurement for studies of exposures.

2. Methods

2.1. Participants

We selected raters (RB, AG, and PR) with master's and doctoral degrees, training in epidemiological methods, and at least four years (range 4-13 years) of experience evaluating epidemiological studies. While the raters did not necessarily have topic-specific expertise on the environmental exposures in the selected systematic reviews, they had access to topic-specific experts and other resources for consultation throughout the project. Raters initially received training materials for each instrument.

2.2. RoB instruments

Raters applied the following instruments by using a specifically prepared Excel package with drop-down response options for each instrument: The RoB instrument for NRS of exposures, OHAT, ORoC, and NOS. In this study we refer to the potential for limitations within a study as 'RoB'; however, some instruments use terms such as 'study quality' or 'study sensitivity' to include or refer to RoB. In addition to the RoB instrument for NRS of exposures, OHAT, ORoC, and NOS are available online and in print [8, 12, 13, 18]. Appendix A presents the domains, domain-, individual study-, and outcome-level responses for each instrument.

2.3. Case-study topics

We used existing systematic reviews on environmental exposures to identify case-study topics (Table 1) [19-24]. We selected the following topics: bisphenol-A (BPA),

perfluorooctanoic acid (PFOA), polybrominated diphenyl ethers (PBDE), particulate matter with aerodynamic diameter less than 2.5 µm ($PM_{2.5}$), folic acid, and phthalate metabolites. We selected a sub-set of 5-to-6 individual studies from each topic to represent a range of study design and exposure assessment features commonly encountered in environmental health. If previous assessments of bias were available in the systematic reviews, then we (RM, KT, HJS) selected individual studies to represent a broad range of potential RoB (from highest to lowest RoB). We considered reports on RoB and results of effect size, confidence interval range, and reported confounders adjusted for in each study's analysis to make a determination on hypothesized RoB.

2.4. Analysis

2.4.1. Interrater and inter-instrument reliability

To evaluate the interrater reliability for study-level assessments for the RoB instrument for NRS of exposures, we calculated an intraclass correlation (ICC) by assessing the five or six studies from the six systematic reviews for a total of 35 studies [19-21, 23-25]. Three raters independently evaluated studies. We also calculated the interrater reliability across the other three instruments using the same 35 individual studies. To protect against order effects, we randomized the order of the four RoB instruments that raters used to assess each of the studies, as well as the order of the studies. We provided one rater with three packages of instrument templates and studies to evaluate and disseminate to the other raters. Raters recorded their responses in Microsoft Excel. Since the four instruments used different study-level RoB judgments (Appendix A), we standardized the distribution of the different ordinal ratings by converting them to z-

scores. Study-level judgments for the RoB instrument for NRS of exposures and ORoC included an option to rate the entire study as either 'No information' (in the RoB instrument for NRS of exposures) or 'Inadequate/Uninformative' (in ORoC). Similarly, these response options could be selected at the domain-level. If raters reported 'Inadequate/Uninformative', we used it as the most severe rating based on the interpretation in the ORoC manual [13]. If raters reported 'No information', we conducted a sensitivity analysis using the most severe and the least severe ratings. Based on raters' feedback during the preliminary application of the instrument, 'No information' may be inappropriately selected when the question may not be intuitive to a rater, but bias may be present (e.g. responding to a question about temporality when evaluating a cross-sectional study). For the statistical analyses, we used SPSS for Windows version 13.0 (SPSS, Chicago, IL, USA).

To understand the degree of contribution of the individual instruments to the variance in the reliability analyses, we also evaluated interrater and inter-instrument reliability. We used generalizability (G) theory and the software program G_String_IV, version 6.1.1 (Hamilton, ON, CA) to estimate reliability and determine the relative contributions of different variance sources (i.e., instrument versus rater) in the data set [26]. These variance components calculate the proportion of error variance (σ) attributed to the object of measurement (individual studies), modelled facet (rater; τ), and the interactions between the object and modelled facets of generalizability (δ or Δ). We used the following formula to calculate the absolute error coefficient (AEC) here: $\mathrm{G} = ((\tau)^2)/\ (\sigma(\tau)^2 + \sigma(\Delta)^2)$). We considered the strength of agreement for the ICCs and AEC

as the following: 0.00 to 0.20 as slight; 0.21 to 0.40 as fair; 0.41 to 0.60 as moderate; 0.61 to 0.80 as substantial; and 0.81 to 1.00 as almost perfect [27].

2.4.2. Construct validity

We assessed construct validity of the RoB instrument for NRS of exposures to determine the extent to which the RoB instrument for NRS of exposures related to other measures consistent with RoB of exposure. In the absence of a reference standard, we conducted a series of correlation analyses with other RoB instruments in the field and topic-specific expert feedback. We used the same 35 studies from the systematic reviews listed above.

Initially, we calculated Pearson correlation coefficients (*r*) between the study-level judgments of the RoB instrument for NRS of exposures and OHAT, ORoC, and NOS by using their average ratings. We hypothesized that the RoB instrument for NRS of exposures would correlate strongly with OHAT, ORoC, and NOS scores since they should measure the same constructs. The degree of correlation between the RoB instrument for NRS of exposures and the other instruments provided insight into whether or not the RoB instrument for NRS of exposures measures RoB within studies of exposure. In addition, we calculated Pearson *r* at the domain level across instruments. We explored scatterplots showing the relationship between each pair of domains to identify potentially spuriously high correlation coefficients, as we would expect linear relations between judgments on similar domains of concepts. We grouped domains of similar concepts (e.g. confounding in the RoB instrument for NRS of exposures versus confounding in OHAT). The NOS lumps questions into three domains: selection,

comparability (i.e. confounding), and outcome; however, when used for case-control studies, NOS refers to the third domain as exposure. Therefore, we compared the domain of outcome twice, once with the group of outcome bias domains and once with the exposure bias domains. In addition, we calculated Pearson *r* across all domains to see if any correlated with concepts that did not seem to overlap (e.g. confounding in the RoB instrument for NRS of exposures vs. selection bias in OHAT), including domains within the same instrument.

Then, we assessed validity by comparing study-level judgments of the instruments against a global rating of RoB from topic-specific experts recognized as authorities in environmental health. We recruited topic-specific experts across five environmental-exposure disciplines (BPA, PFOA, PBDE, $PM_{2.5}$, and phthalates) for the evaluation of 29 studies (Appendix B). Twenty-eight PhD-level topic-specific experts provided 160 observations of unstructured RoB judgments utilizing a 7-point Likert-scale to express agreement with the following statement: 'The study that you just reviewed is of low risk of bias.' (Figure 1). In addition, all topic-specific experts rated the overall RoB for each study on a four-point scale from 'Low risk of bias' to 'Critical risk of bias.' Topic-specific experts reviewed the papers independently of any rating instruments. We considered correlation coefficients of 0.10 to 0.30 as weak, 0.30 to 0.60 as moderate, and > 0.60 as strong [28].

2.4.3. Comparison of RoB usability across the instruments

We documented the time required to complete ratings as a surrogate to evaluate the burden of using the instruments. We calculated the mean time in minutes for the three

raters who evaluated the studies across the four instruments and provided the range. In addition, the external topic-specific experts reported time to review each study. When a range of time was given, the most conservative (longest) estimate of time was used in the analysis. In SPSS, we performed an analysis of variance (ANOVA) to analyze the difference between the mean duration reported by the raters and topic-specific experts followed by Tukey's *post hoc* test to explore significant differences between each instrument and topic experts.

2.4.4. Sample size estimation

We used the individual study-level RoB judgment to determine the sample size. We determined *a priori* that a standard error of ± 0.10 provided sufficient precision for a ICC of 0.75 [26] and required five-to-six studies per review for a total of 35 individual studies to assess the instruments.

3. Results

3.1. Interrater and inter-instrument reliability of RoB instruments

Interrater reliability of individual study-level judgments when using the RoB instrument for NRS of exposures demonstrated substantial agreement (ICC = 0.73; 95% CI: 0.53, 0.85). OHAT, ORoC, and NOS showed similar substantial interrater agreement (Table 2).

The interrater and inter-instrument reliability of the individual study-level judgments demonstrated substantial generalizability (AEC = 0.70) to another set of raters or instruments. Further exploration of the individual interrater reliability estimates suggests that the consistency among raters' responses is greater than the AEC. Since the

AEC aggregates reliability from the interrater and the inter-instrument estimates, this suggests that the different instruments introduce some variance.

3.2. Validity of RoB instruments

3.2.1. Between instrument total score correlations

Table 3 presents the average measures of correlation coefficients between the instruments and topic-specific experts at the study level. We observed strong correlations between the RoB instrument for NRS of exposures and the other RoB instruments (r = 0.74 to 0.90) suggesting agreement across instruments for the study-level judgments (i.e. the study-level judgments are similar across instruments). In absence of a reference standard to measure RoB of exposure studies, the strength of the correlations suggests that all instruments are measuring similar concepts of bias at the study level. The strongest agreement existed between the RoB instrument for NRS of exposures and the OHAT instrument (r = 0.90; 95% CI: 0.81, 0.95). We found somewhat weaker correlations between the RoB instrument for NRS of exposures and the OHAT tool with the NOS and the topic experts' global judgment.

3.2.2. Between instrument related domain score correlations

Correlations at the domain level suggest greater variability in the measurement of specific RoB concepts compared to the instrument total scores. Comparisons between similar domains across the instruments suggest that validity differs across the instruments and domains (Appendix C).

The RoB instrument for NRS of exposures showed a strong correlation (r = 0.69) with OHAT and NOS (r = 0.79) for the domain of confounding. The RoB instrument for NRS of exposures, OHAT, and ORoC demonstrated strong correlations across the domain of exposure measurement ($r \geq 0.92$) suggesting near identical measurement of the concept. Whereas, the weak correlation of the comparable concept in the NOS suggests that the NOS is not measuring risk of bias due to exposure misclassification well. Only the RoB instrument for NRS of exposures and ORoC presented judgments on departures from intended exposures. The correlation was only moderate (r = 0.49). The RoB instrument for NRS of exposures and OHAT demonstrated a strong correlation in the measurement of missing data (r = 0.61); however, the RoB instrument for NRS of exposures and ORoC showed lower correlations. For the domain of measurement of outcomes, the RoB instrument for NRS of exposures, OHAT, and ORoC demonstrated moderate to strong correlations of the concept (r = 0.52 to 0.63); however, comparison with NOS revealed weak correlations. Correlations between the RoB instrument for NRS of exposures, OHAT, and ORoC were weak for the concept of reporting bias.

3.2.3. Within instrument domain score correlations

Correlation coefficients between all domains within the same instrument suggest the measurement of overlapping concepts for many but not all domains (Appendix D). High correlation coefficients suggest that the risk of bias domains measure similar concepts of bias occurring together in the studies (e.g. selection bias and measurement of exposure bias occur together). For the RoB instrument for NRS of exposures, correlation coefficients were moderate for bias due to selection of participants and bias due to

confounding (*r* = 0.51), bias due to intended exposures (*r* = 0.64), and bias in measurement of outcomes (*r* = 0.63).

The within instrument domain score correlations were also moderate to strong for the other instruments. For the OHAT instrument, correlation coefficients demonstrated the measurement of similar concepts between the domain of selection bias and confounding bias (*r* = 0.51), detection bias (*r* = 0.54), selective reporting bias (*r* = 0.52), and other sources of bias (*r* = 0.58). In addition, a moderate correlation existed between confounding bias and selective reporting bias (*r* = 0.53), as well as other sources of bias (*r* = 0.51); outcome bias demonstrated moderate correlations with selective reporting bias (*r* = 0.51) and other sources of bias (*r* = 0.50). For the ORoC, selection bias showed moderate correlations with confounding (r = 0.50), selective reporting (r = 0.53), and analysis (r = 0.59). High correlations were present between outcome and analysis (*r* = 0.77), as well as study sensitivity (*r* = 0.61). In addition, correlations between analysis, confounding (*r* = 0.57), and study sensitivity (*r* = 0.66) were moderate to strong. Of the three domains within NOS, selection bias and comparability demonstrated a moderate correlation (*r* = 0.57).

3.2.4. Between instrument domain score correlations

Appendix E presents moderate and strong Pearson correlation coefficients of unrelated domains across different instruments. The bias due to confounding domain in RoB instrument for NRS of exposures correlated strongly with the selection bias domains in both OHAT and ORoC instruments (*r* = 0.69 and *r* = 0.64, respectively), and the analysis domain in ORoC (*r* = 0.71). Bias due to missing data demonstrated a strong correlation

with ORoC reporting (r = 0.79). Correlation coefficients were moderate to high between the majority of OHAT and ORoC domains and unrelated domains in other instruments, except for the domains evaluating the exposure. Specifically, correlation coefficients for the OHAT domain of selection bias showed higher correlations with the RoB instrument for NRS of exposures domain of bias due to confounding (r = 0.67) and NOS comparability (r = 0.53), than correlations with the respective selection bias domains in those instruments. Other sources of bias in the OHAT instrument were moderately to strongly correlated with RoB instrument for NRS of exposures missing data (r = 0.52); and ORoC selection bias (r = 0.54), outcome (r = 0.73), reporting (r = 0.57), and analysis (r = 0.67). The NOS domain of comparability correlated moderately with several different domains in the ORoC: selection (r = 0.60), analysis (r = 0.56), and study sensitivity (r = 0.54).

3.3. Instrument burden

Raters and topic-specific experts provided the approximate time per individual study evaluated. Across the four instruments evaluated by raters, time per study varied between 5 and 150 minutes. The mean (range) time estimate in minutes for applying each instrument to a single study is as follows: RoB instrument for NRS of exposures: 79 (30 – 150); OHAT: 39 (15 – 60); ORoC: 31 (10 – 60); and NOS: 12 (5 – 30). Topic-specific experts reported a mean (range) in minutes for their unstructured evaluations of studies: 42 (15 – 150). Further analyses demonstrated significant differences between the time estimates reported for each instrument and time estimates reported from topic-specific experts (Appendix F). The RoB instrument of exposures required

significantly more time than all other instruments ($p < 0.001$), and the NOS required significantly less time than all other instruments ($p < 0.01$ for all comparisons).

4. Discussion

4.1. Statement of principle findings

Our study results indicate that the RoB instrument for NRS of exposures is a reliable and valid instrument for assessing RoB in studies of exposures. The interrater and inter-instrument analyses suggest that these results will be generalizable to the application of these instruments beyond the raters used in our study. The observed construct validity suggests that the RoB instrument for NRS of exposures is accurately measuring the concept of bias in studies of exposures. We believe that the somewhat lower correlations between the RoB instrument for NRS of exposures and the OHAT with the NOS and global judgments suggest better identification of risk of bias in various domains.

While the comparison of time to complete each of the RoB assessments varied significantly across instruments and the topic-specific experts, suggesting the greatest burden from the RoB instrument for NRS of exposures, there are several other considerations for selecting the most appropriate and efficient instrument to use. These considerations include the measurement of relevant concepts to determine the RoB in studies of exposures, exploration of the overlap of concepts within instruments, the detail of the signaling questions, and the transparency of concepts considered in the RoB judgment.

This study adds important information about the measurement properties of risk of bias instruments for exposure studies. First, the measurement of the exposure is essential in any RoB instrument for studies of exposures. While one version of the NOS labels a domain as exposure [18], the lack of agreement with the RoB instrument for studies of exposures, OHAT, and ORoC and the substantial agreement between that domain among those latter three instruments suggests that the NOS does not identify risk of bias in this domain well. The correlation coefficients suggest minimal differences between the RoB instrument for NRS of exposures, OHAT, and ORoC on this domain; therefore, one could select that domain from any of those instruments. The domains that measure the concept of exposure in the RoB instrument for NRS of exposures, OHAT and ORoC do not correlate highly with any unrelated domain. This finding suggests the importance of separate assessment of this domain and that there is little overlap in bias or assessment with any of the other domains.

Second, the assessment of the validity by domain reveals important observations. For the domains of confounding and participant selection, the correlations within instruments suggest that further exploration is needed to either distinguish between the two concepts or to merge them, as currently the same concepts may be measured multiple times within the same instrument. The RoB instrument for NRS of exposures domains remain conceptually somewhat distinct, as suggested by the mostly moderate correlation coefficients between domains in the instrument. It appears that this instrument showed generally weaker correlations with unrelated domains in other instruments, while OHAT, ORoC, and NOS domains demonstrated more correlations

with several unrelated domains in other instruments. Over a third of domains within OHAT, ORoC, and NOS suggest moderate or substantial correlations with other domains in the same instrument, suggesting frequent overlap of the same concepts. For example, the OHAT domain of other sources of bias demonstrates high correlation with several distinct domains on the ORoC instrument (selection bias, outcome, reporting, and analysis), even though the OHAT instrument contains domains with these labels. These somewhat higher correlations suggest that OHAT, ORoC, and NOS might benefit from further scale development, such as factor analysis of individual items and validity testing of domains to identify grouping of items to form discrete domains.

Third, while the required time to apply the RoB instrument for NRS of exposures is significantly greater than for all other instruments and the global judgment by topic-specific experts, the level of detail in the signaling questions may allow users to make more informed decisions. However, this burden may be important when performing large scale systematic reviews that require assessment of many studies with possibly limited gain in accuracy of RoB assessment.

Fourth, including study-level judgments from topic-specific experts allowed us to observe the level of agreement between judgments of bias made implicitly by experts and compare those to RoB instruments that aim to make judgments explicit. We identified strong agreement between study-level judgments from the RoB instrument for NRS of exposures and the topic-specific experts' evaluations. This instrument serves as a reminder of important concepts and may be just as good for application by persons who are not as familiar with studies of exposures or specific types of exposures.

4.2. Strengths and limitations

This is the first study to present evidence of the measurement properties of instruments to evaluate RoB in studies of exposure, in particular environmental exposures. While the measurement properties of the RoB instrument for NRS of exposures were our primary focus, we evaluated three additional instruments. The large sample of studies and across six distinct exposures and health outcomes to broaden the robustness and generalizability of our results represent a strength. We calculated correlation coefficients and conducted a G theory analysis. The G theory AEC demonstrates substantial generalizability to allow us to say that these results would also be generated by another set of raters or across instruments. It is further confirmed by the individual ICC and correlation coefficients. In addition, our analyses of each domain across all four instruments provides greater understanding of the individual concepts used to measure studies of exposures. Lastly, when conducting the domain-level analysis, we recognized that instruments include unique but potentially similar domain concepts and groupings of signaling questions (e.g., 'bias in missing data' in the RoB instrument for studies of exposure and 'attrition' in OHAT); therefore, we were able to examine correlation coefficients to assure that we captured domains measuring similar concepts.

We recognize some limitations within our study. First, raters evaluated the same studies multiple times with different instruments. To account for potential order bias during the assessment process, we randomized the order of the four instruments and the studies for each instrument. Second, topic-specific experts provided evaluations on all topics except for folic acid. This was due to difficulties with recruitment of experts. We do not

believe that this had an impact on the feedback that we received on the instrument, as topic-specific experts provided 160 observations across the other five exposure topics. However, this could provide a less generalizable response as fewer topic areas were assessed. Third, each instrument provided different options for rating RoB, including options that could be interpreted as more or less bias, such as 'Inadequate' or 'No information.' To account for the different RoB judgment options, we normalized the results to allow the calculation of correlation coefficients. However, standardizing the results cannot take into account the differences between the number of signaling questions and response options, which vary between all instruments. As demonstrated by the generalizability analysis, the instruments introduced some variance. This could be due to the limited response options and prescriptive nature of NOS versus the numerous response options and subjective judgments used by the other instruments. In addition, no study-level RoB judgment included categorical responses (e.g., 'Inadequate'). Fourth, our domain-level analysis compared the conceptual groupings provided in each instrument. We did not perform an item-level or factor analysis of the signaling questions grouped within each domain. Therefore, correlation coefficients may suggest less agreement from domains that contained questions addressing multiple concepts. For example, in the NOS outcome domain, the three questions ask about the adequacy of the assessment of the outcome, the duration of follow up, and missing data. In other instruments, such as the RoB instrument for NRS of exposures and OHAT, missing data or attrition is a separate domain. Comparing concepts at the item level may produce higher correlations (e.g., the NOS question about missing data compared with

the RoB instrument for NRS of exposures domain on bias in missing data); however, as explained previously, the granularity and order of the multiple signaling questions of the RoB instrument for NRS of exposures may help users make more informed decisions about bias and should be explored. Fifth, we did not identify a reference standard of RoB of exposure studies, which would have provided an ideal comparison to measure validity. Instead, to explore the validity of the RoB instrument for NRS of exposures, we compared it with other established RoB instruments for studies of exposures. In addition, we identified several topic-specific experts to review individual studies for the construct validity analyses. Lastly, we did not calculate test-retest reliability. While the generalizability coefficient suggests that another set of raters would be able to use these instruments with substantial agreement, we cannot address the stability of the instruments' scores over time.

4.3. Implications for researchers and policymakers

The RoB instrument for NRS of exposures provides a reliable and validated instrument for systematic review developers. While we recognize the time implications of applying this instrument, there are several benefits to consider, specifically the RoB instrument for NRS of exposures measures distinct concepts in each domain and validly measures the potential for bias from exposures.

RoB judgments across a body of evidence for a given outcome can help guideline developers and policy makers evaluate and interpret the certainty in a body of evidence to inform decision making. When using the RoB instrument for NRS of exposures, systematic review authors and guideline developers can start the GRADE evidence

assessment at 'high' initial certainty; whereas OHAT, ORoC, and NOS lack the standardized evaluation of NRS to a randomized experiment and use a study-design driven approach for evaluation.

4.4. Unanswered questions and future research

Future research should examine the precision with which each of these instruments discriminates between measurement domains. Instruments developed for wider evaluation across subjects may have less ability to detect small changes in attributes when compared with instruments developed explicitly for studies of environmental exposures. Our use of G-theory, as well as ICC and correlation coefficient analyses provides guidance facilitating future studies evaluating the measurement properties of instruments. Due to the plethora of RoB instruments available for NRS of health interventions, a similarly structured study would produce valuable information for methods advancement.

5. Conclusions

The RoB instrument for NRS of exposures is ready for application by a wider audience. This article presents several variables for consideration when deciding on an instrument to use to evaluate RoB of studies of environmental exposure, such as reliability across users and instruments, organization of concepts within instruments, and expected time expenditure. The RoB instrument for NRS of exposures has comparable reliability and validity to other instruments used for evaluating RoB in exposure studies; however, this instrument suggests increased efficiency of construct measurement and acuity when

evaluating the potential for bias from studies of exposures. Based on the standardized approach to compare a study against an unbiased hypothetical study, the RoB instrument for NRS of exposures is the only instrument allowing systematic review and guideline developers to start at 'high' initial certainty in GRADE, although risk of bias will usually lead to rating down by at least two levels.

6. Acknowledgments

We would like to acknowledge Geoff Norman for his contributions to the

conceptualization of the study design.

7. Funding Sources

This research was supported by the Intramural Research Program of the National Institute of Environmental Health Sciences and the GRADE Centre at the McMaster University.

8. References

1. Shamliyan T, Kane RL, Dickinson S: **A systematic review of tools used to assess the quality of observational studies that examine incidence or prevalence and risk factors for diseases**. *J Clin Epidemiol* 2010, **63**(10):1061-1070.
2. Deeks JJ, Dinnes J, D'amico R, Sowden A, Sakarovitch C, Song F, Petticrew M, Altman D: **Evaluating non-randomised intervention studies**. *Health technology assessment* 2003, **7**(27):1-179.
3. **Cochrane Handbook for Systematic Reviews of Interventions** [www.cochrane-handbook.org]
4. Voss PH, Rehfuess EA: **Quality appraisal in systematic reviews of public health interventions: an empirical study on the impact of choice of tool on meta-analysis**. *J Epidemiol Community Health* 2013, **67**(1):98-104.
5. Rooney AA, Cooper GS, Jahnke GD, Lam J, Morgan RL, Boyles AL, Ratcliffe JM, Kraft AD, Schünemann HJ, Schwingl P: **How credible are the study results? Evaluating and applying internal validity tools to literature-based assessments of environmental health hazards**. *Environment international* 2016.
6. Morgan RL, Thayer KA, Bero L, Bruce N, Falck-Ytter Y, Ghersi D, Guyatt G, Hooijmans C, Langendam M, Mandrioli D *et al*: **GRADE: Assessing the quality of evidence in environmental and occupational health**. *Environ Int* 2016, **92-93**:611-616.
7. Schünemann H, Cuello C, Akl EA, Mustafa R, Meerpohl J, Thayer K, Morgan RL, Gartlehner G, Kunz R, Katikireddi S *et al*: **GRADE Guidelines: 18. How tools to assess risk of bias in non-randomized studies should be used to rate the certainty of a body of evidence** *Journal of clinical epidemiology* Accepted for publication.
8. Morgan RL, Thayer K, Santesso N, Holloway AC, Blain R, Eftim S, Goldstone A, Ross P, Guyatt G, Schünemann H: **Need for an instrument to evaluate Risk of Bias in Non-randomized Studies of Exposure: Rationale and preliminary instrument**. *Environment International* Under review.
9. **ROBINS-I: a tool for assessing Risk Of Bias In Non-randomized Studies of Interventions**
10. Guyatt GH: **Measuring health-related quality of life: General issues**. *Canadian Respiratory Journal* 1997, **4**(3):123-130.
11. NRC (National Research Council): **Review of EPA's Integrated Risk Information System (IRIS) Process (http://www.nap.edu/catalog.php?record_id=18764) [accessed 1 January 2015]**. 2014.
12. NTP (National Toxicology Program): **Handbook for Conducting a Literature-Based Health Assessment Using Office of Health Assessment and Translation (OHAT) Approach for Systematic Review and Evidence Integration. January 9, 2015 release. Available at http://ntp.niehs.nih.gov/go/38673.** 2015.

13. NTP (National Toxicology Program): **Handbook for Preparing Report on Carcinogens Monographs - July 2015. Available at http://ntp.niehs.nih.gov/go/rochandbook**. 2015(January 3, 2017).
14. Woodruff TJ, Sutton P: **The Navigation Guide systematic review methodology: a rigorous and transparent method for translating environmental health science into better health outcomes**. *Environ Health Perspect* 2014, **122**(10):1007-1014.
15. Cooper GS, Lunn RM, Agerstrand M, Glenn BS, Kraft AD, Luke AM, Ratcliffe JM: **Study sensitivity: Evaluating the ability to detect effects in systematic reviews of chemical exposures**. *Environ Int* 2016, **92-93**:605-610.
16. Lin CK, Hung HY, Christiani DC, Forastiere F, Lin RT: **Lung cancer mortality of residents living near petrochemical industrial complexes: a meta-analysis**. *Environ Health* 2017, **16**(1):101.
17. Lewis-Mikhael AM, Bueno-Cavanillas A, Ofir Guiron T, Olmedo-Requena R, Delgado-Rodriguez M, Jimenez-Moleon JJ: **Occupational exposure to pesticides and prostate cancer: a systematic review and meta-analysis**. *Occup Environ Med* 2016, **73**(2):134-144.
18. Wells G, Shea B, O'connell D, Peterson J, Welch V, Losos M, Tugwell P: **The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. Ottawa Hospital Research Institute, 2014**. In.: oxford.asp; 2015.
19. Ferguson KK, O'Neill MS, Meeker JD: **Environmental contaminant exposures and preterm birth: a comprehensive review**. *J Toxicol Environ Health B Crit Rev* 2013, **16**(2):69-113.
20. Hamra GB, Guha N, Cohen A, Laden F, Raaschou-Nielsen O, Samet JM, Vineis P, Forastiere F, Saldiva P, Yorifuji T *et al*: **Outdoor particulate matter exposure and lung cancer: a systematic review and meta-analysis**. *Environ Health Perspect* 2014, **122**(9):906-911.
21. Johnson PI, Sutton P, Atchley DS, Koustas E, Lam J, Sen S, Robinson KA, Axelrad DA, Woodruff TJ: **The Navigation Guide - evidence-based medicine meets environmental health: systematic review of human evidence for PFOA effects on fetal growth**. *Environ Health Perspect* 2014, **122**(10):1028-1039.
22. Muggli EE, Halliday JL: **Folic acid and risk of twinning: a systematic review of the recent literature, July 1994 to July 2006**. *Med J Aust* 2007, **186**(5):243-248.
23. Ranciere F, Lyons JG, Loh VH, Botton J, Galloway T, Wang T, Shaw JE, Magliano DJ: **Bisphenol A and the risk of cardiometabolic disorders: a systematic review with meta-analysis of the epidemiological evidence**. *Environ Health* 2015, **14**(1):46.
24. Zhao X, Wang H, Li J, Shan Z, Teng W, Teng X: **The Correlation between Polybrominated Diphenyl Ethers (PBDEs) and Thyroid Hormones in the General Population: A Meta-Analysis**. *PLoS One* 2015, **10**(5):e0126989.
25. Boyles AL, Yetley EA, Thayer KA, Coates PM: **Safe use of high intakes of folic acid: research challenges and paths forward**. *Nutrition reviews* 2016, **74**(7):469-474.

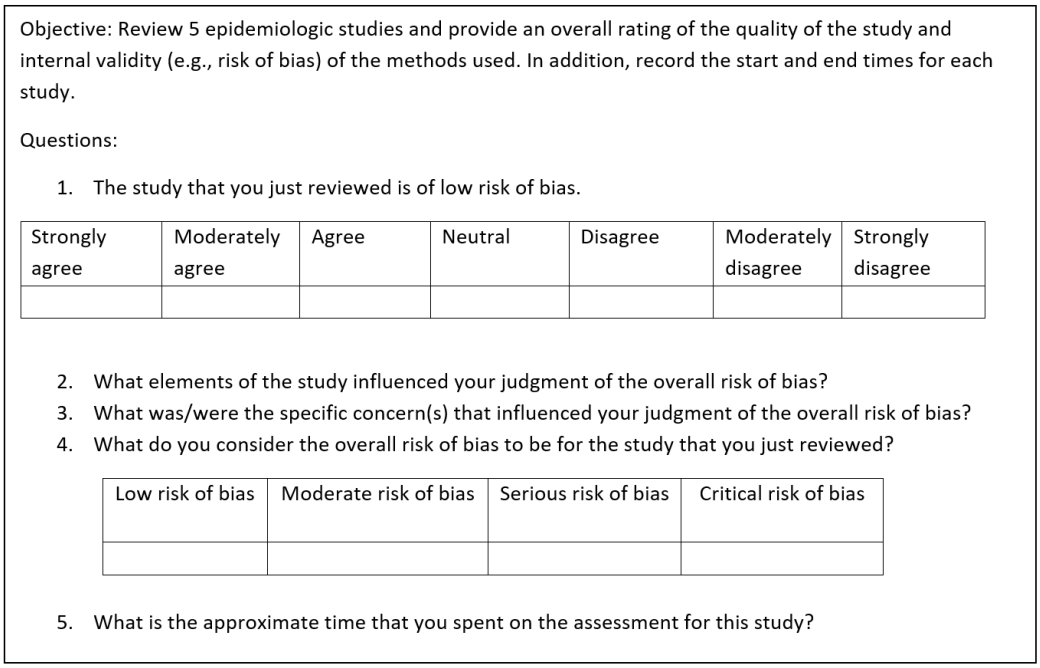
26. Streiner DL, Norman GR, Cairney J: **Health measurement scales: a practical guide to their development and use**: Oxford university press; 2014.
27. Landis JR, Koch GG: **The measurement of observer agreement for categorical data**. *Biometrics* 1977, **33**(1):159-174.
28. Cohen J: **Statistical power analysis for the behavioral sciences (revised ed.)**. In.: New York: Academic Press; 1977.
29. Apelberg BJ, Witter FR, Herbstman JB, Calafat AM, Halden RU, Needham LL, Goldman LR: **Cord serum concentrations of perfluorooctane sulfonate (PFOS) and perfluorooctanoate (PFOA) in relation to weight and size at birth**. *Environmental health perspectives* 2007:1670-1676.
30. Hamm MP, Cherry NM, Chan E, Martin JW, Burstyn I: **Maternal exposure to perfluorinated acids and fetal growth**. *Journal of Exposure Science and Environmental Epidemiology* 2010, **20**(7):589-597.
31. Kim S-K, Lee KT, Kang CS, Tao L, Kannan K, Kim K-R, Kim C-K, Lee JS, Park PS, Yoo YW: **Distribution of perfluorochemicals between sera and milk from the same mothers and implications for prenatal and postnatal exposures**. *Environmental Pollution* 2011, **159**(1):169-174.
32. Maisonet M, Terrell ML, McGeehin MA, Christensen KY, Holmes A, Calafat AM, Marcus M: **Maternal concentrations of polyfluoroalkyl compounds during pregnancy and fetal and postnatal growth in British girls**. *Environmental health perspectives* 2012, **120**(10):1432-1437.
33. Nolan LA, Nolan JM, Shofer FS, Rodway NV, Emmett EA: **The relationship between birth weight, gestational age and perfluorooctanoic acid (PFOA)-contaminated public drinking water**. *Reproductive Toxicology* 2009, **27**(3):231-238.
34. Whitworth KW, Haug LS, Baird DD, Becher G, Hoppin JA, Skjaerven R, Thomsen C, Eggesbo M, Travlos G, Wilson R: **Perfluorinated compounds in relation to birth weight in the Norwegian Mother and Child Cohort Study**. *American journal of epidemiology* 2012:kwr459.
35. Bhandari R, Xiao J, Shankar A: **Urinary bisphenol A and obesity in US children**. *American journal of epidemiology* 2013, **177**(11):1263-1270.
36. Carwile JL, Michels KB: **Urinary bisphenol A and obesity: NHANES 2003-2006**. *Environ Res* 2011, **111**(6):825-830.
37. Harley KG, Aguilar Schall R, Chevrier J, Tyler K, Aguirre H, Bradman A, Holland NT, Lustig RH, Calafat AM, Eskenazi B: **Prenatal and postnatal bisphenol A exposure and body mass index in childhood in the CHAMACOS cohort**. *Environmental health perspectives* 2013, **121**(4):514-520.
38. Shankar A, Teppala S, Sabanayagam C: **Urinary bisphenol A levels and measures of obesity: results from the national health and nutrition examination survey 2003–2008**. *ISRN endocrinology* 2012, **2012**.
39. Wang H, Zhou Y, Tang C, Wu J, Chen Y, Jiang Q: **Association between bisphenol A exposure and body mass index in Chinese school children: a cross-sectional study**. *Environ Health* 2012, **11**(1):79.

40. Zhao HY, Bi YF, Ma LY, Zhao L, Wang TG, Zhang LZ, Tao B, Sun LH, Zhao YJ, Wang WQ *et al*: **The effects of bisphenol A (BPA) exposure on fat mass and serum leptin concentrations have no impact on bone mineral densities in non-obese premenopausal women**. *Clinical Biochemistry* 2012, **45**(18):1602-1606.
41. Hystad P, Demers PA, Johnson KC, Carpiano RM, Brauer M: **Long-term residential exposure to air pollution and lung cancer risk**. *Epidemiology* 2013, **24**(5):762-772.
42. Cao J, Yang C, Li J, Chen R, Chen B, Gu D, Kan H: **Association between long-term exposure to outdoor air pollution and mortality in China: a cohort study**. *Journal of hazardous materials* 2011, **186**(2):1594-1600.
43. Katanoda K, Sobue T, Satoh H, Tajima K, Suzuki T, Nakatsuka H, Takezaki T, Nakayama T, Nitta H, Tanabe K *et al*: **An Association Between Long-Term Exposure to Ambient Air Pollution and Mortality From Lung Cancer and Respiratory Diseases in Japan**. *Journal of Epidemiology* 2011, **21**(2):132-143.
44. Lepeule J, Laden F, Dockery D, Schwartz J: **Chronic exposure to fine particles and mortality: an extended follow-up of the Harvard Six Cities study from 1974 to 2009**. *Environ Health Perspect* 2012, **120**(7):965-970.
45. Cesaroni G, Badaloni C, Gariazzo C, Stafoggia M, Sozzi R, Davoli M, Forastiere F: **Long-Term Exposure to Urban Air Pollution and Mortality in a Cohort of More than a Million Adults in Rome**. *Environmental Health Perspectives* 2013, **121**(3):324-331.
46. Krewski D, Jerrett M, Burnett RT, Ma R, Hughes E, Shi Y, Turner MC, Pope III CA, Thurston G, Calle EE: **Extended follow-up and spatial analysis of the American Cancer Society study linking particulate air pollution and mortality**: Health Effects Institute Boston, MA; 2009.
47. Bloom M, Spliethoff H, Vena J, Shaver S, Addink R, Eadon G: **Environmental exposure to PBDEs and thyroid function among New York anglers**. *Environmental toxicology and pharmacology* 2008, **25**(3):386-392.
48. Lin SM, Chen FA, Huang YF, Hsing LL, Chen LL, Wu LS, Liu TS, Chang-Chien GP, Chen KC, Cho HR: **Negative associations between PBDE levels and thyroid hormones in cord blood**. *International Journal of Hygiene and Environmental Health* 2011, **214**(2):115-120.
49. Stapleton HM, Eagle S, Anthopolos R, Wolkin A, Miranda ML: **Associations between Polybrominated Diphenyl Ether (PBDE) Flame Retardants, Phenolic Metabolites, and Thyroid Hormones during Pregnancy**. *Environmental Health Perspectives* 2011, **119**(10):1454-1459.
50. GuanGen H, GangQiang D, XiaoMing L, XiaoFeng W, JianLong H, HaiTao S, Yu Z, LeYan D: **Correlations of PCBs, DIOXIN, and PBDE with TSH in children's blood in areas of computer E-waste recycling**. *Biomedical and Environmental Sciences* 2011, **24**(2):112-116.
51. Kim TH, Bang du Y, Lim HJ, Won AJ, Ahn MY, Patra N, Chung KK, Kwack SJ, Park KL, Han SY *et al*: **Comparisons of polybrominated diphenyl ethers levels in paired South Korean cord blood, maternal blood, and breast milk samples**. *Chemosphere* 2012, **87**(1):97-104.

52. Kim S, Park J, Kim HJ, Lee JJ, Choi G, Choi S, Kim S, Kim SY, Moon HB, Kim S *et al*: **Association between several persistent organic pollutants and thyroid hormone levels in serum among the pregnant women of Korea**. *Environment International* 2013, **59**:442-448.
53. Adibi JJ, Hauser R, Williams PL, Whyatt RM, Calafat AM, Nelson H, Herrick R, Swan SH: **Maternal urinary metabolites of di-(2-ethylhexyl) phthalate in relation to the timing of labor in a US multicenter pregnancy cohort study**. *American journal of epidemiology* 2009, **169**(8):1015-1024.
54. Suzuki Y, Niwa M, Yoshinaga J, Mizumoto Y, Serizawa S, Shiraishi H: **Prenatal exposure to phthalate esters and PAHs and birth outcomes**. *Environment international* 2010, **36**(7):699-704.
55. Whyatt RM, Adibi JJ, Calafat AM, Camann DE, Rauh V, Bhat HK, Perera FP, Andrews H, Just AC, Hoepner L: **Prenatal di (2-ethylhexyl) phthalate exposure and length of gestation among an inner-city cohort**. *Pediatrics* 2009, **124**(6):e1213-e1220.
56. Wolff MS, Engel SM, Berkowitz GS, Ye X, Silva MJ, Zhu C, Wetmur J, Calafat AM: **Prenatal phenol and phthalate exposures and birth outcomes**. *Environmental health perspectives* 2008, **116**(8):1092-1097.
57. Meeker JD, Hu H, Cantonwine DE, Lamadrid-Figueroa H, Calafat AM, Loch-Caruso R, Téllez-Rojo MM, Ettinger AS, Hernandez-Avila M: **Urinary phthalate metabolites in relation to preterm birth in Mexico City**. 2009.

Figures

Figure 1.



Objective: Review 5 epidemiologic studies and provide an overall rating of the quality of the study and internal validity (e.g., risk of bias) of the methods used. In addition, record the start and end times for each study.

Questions:

1. The study that you just reviewed is of low risk of bias.

Strongly agree	Moderately agree	Agree	Neutral	Disagree	Moderately disagree	Strongly disagree

2. What elements of the study influenced your judgment of the overall risk of bias?
3. What was/were the specific concern(s) that influenced your judgment of the overall risk of bias?
4. What do you consider the overall risk of bias to be for the study that you just reviewed?

Low risk of bias	Moderate risk of bias	Serious risk of bias	Critical risk of bias

5. What is the approximate time that you spent on the assessment for this study?

Figure 1. Survey questions to measure study quality and RoB for topic-specific experts not using a formal RoB instrument.

Tables

Table 1.

Topic	Source title (reference)	Exposure	Outcome	Systematic review studies (N)	Studies used in current analysis (n)
Perfluorooctanoic acid (PFOA) and birthweight	The Navigation Guide-Evidence-Based Medicine Meets Environmental Health: Systematic Review of Human Evidence for PFOA Effects on Fetal Growth. [21]	PFOA	Fetal growth (i.e., birth weight)	17	6 (Apelberg et al., 2007; Hamm et al., 2010; Kim et al., 2011; Maisonet et al., 2012; Nolan et al., 2009; Whitworth et al., 2012 [29-34])
Bisphenol A (BPA) and overweight and obesity	Bisphenol A and the risk of cardiometabolic disorders: a systematic review with meta-analysis of the epidemiological evidence [23]	BPA	Overweight & obesity	14	6 (Bhandari et al., 2013; Carwile & Michels, 2011; Harley et al., 2013; Shankar et al., 2012; Wang et al., 2012; Zhao et al., 2012 [35-40])
Particulate matter less than 2.5 µm ($PM_{2.5}$) and lung cancer	Outdoor particulate matter exposure and lung cancer: a systematic review and meta-analysis [20]	$PM_{2.5}$	Lung cancer	12	6 (Cao et al., 2011; Cesaroni et al., 2013; Hystad et al., 2013; Katanoda et al., 2011; Krewski et al., 2009; Lepeule et al., 2012 [41-46])
Polybrominated diphenyl ethers (PBDEs) and thyroid	The Correlation between Polybrominated Diphenyl Ethers	PBDE	Thyroid function as measured by thyroid	10	6 (Bloom et al., 2008; Han et al., 2010; Kim et al., 2012; Kim et

stimulating hormones	(PBDEs) and Thyroid Hormones in the General Population: A Meta-Analysis. [24]		simulation hormones (TSHs)		al., 2013; Lin et al., 2011; Stapleton et al., 2011 [47-52])
Folic acid supplementation and twin live births	Folic acid and risk of twinning: a systematic review of the recent literature, July 1994 to July 2006 [22]	Folic acid	Twin live births	9	6 (Ballas, Baxter, and Riddick 2006; Ericson, Källén, and Aberg 2001; Kucik and Correa 2004; Li et al., 2003; Signore et al., 2005; Waller et al., 2003)
Phthalate metabolites and preterm birth	Environmental Contaminant Exposures and Preterm Birth: A Comprehensive Review [19]	Phthalate metabolites	Preterm birth (≤ 37 weeks of gestation)	5	5 (Adibi et al., 2009; Meeker et al., 2009; Suzuki et al., 2010; Whyatt et al., 2009; Wolff et al., 2008 [53-57])

BPA: bisphenol-A; PBDE: polybrominated diphenyl ethers; PFOA: perfluorooctanoic acid; $PM_{2.5}$: particulate matter with aerodynamic diameter less than 2.5 µm; TSH: thyroid simulation hormones

Table 1. Case-study topics and studies selected for analysis.

Table 2.

Instrument	Interrater reliability	AEC
RoB instrument for NRS of exposures	0.73 (95% CI: 0.53, 0.85)	0.70
OHAT	0.80 (95% CI: 0.64, 0.89)	
ORoC	0.70 (95% CI: 0.47, 0.84)	
NOS	0.89 (95% CI: 0.80, 0.94)	

AEC: absolute error coefficient; NOS: Newcastle-Ottawa Scale; NRS: non-randomized studies; OHAT: Office of Health Assessment and Technology; ORoC: Office on the Report on Carcinogens; RoB: risk of bias.

Table 2. Interrater reliability for each individual RoB instrument for studies of exposures and an aggregate interrater and inter-instrument reliability across all instruments.

Table 3.

Instrument	RoB instrument for NRS of exposures	OHAT	ORoC	NOS	Topic-specific experts*
RoB instrument for NRS of exposures	1	0.90 (95% CI: 0.81, 0.95)	0.85 (0.70, 0.92)	0.74 (0.49, 0.87)	0.71 (0.38, 0.86)
OHAT		1	0.89 (0.77, 0.94)	0.77 (0.53, 0.88)	0.76 (0.48, 0.89)
ORoC			1	0.80 (0.61, 0.90)	0.83 (0.65, 0.92)
NOS				1	0.81 (0.59, 0.91)

NOS: Newcastle-Ottawa Scale; OHAT: Office of Health Assessment and Technology; ORoC: Office on the Report on Carcinogens; ROBINS: Risk of Bias for Non-randomized Studies of Exposures. *Based on 160 observations across 29 studies.

Table 3. Average measures correlation coefficients between instruments and topic-specific experts at the study level.

Appendices

Appendix A. Characteristics of four RoB instruments.

	RoB instrument for NRS of exposures[1]	OHAT[2*]	ORoC[3±]	NOS[4]
Domains or items	• Bias due to confounding • Bias in selection of participants into the study • Bias in classification of exposures • Bias due to departures from intended exposures • Bias due to missing data • Bias in measurement of outcomes • Bias in selection of the reported result	• Selection bias • Confounding bias • Performance bias • Detection bias (exposure and outcome) • Selective reporting bias • Other sources of bias	• Exposure • Outcome • Selection bias • Confounding • Analysis and reporting	• Selection • Comparability • Ascertainment of outcome/exposure
Domain/ item-level responses	• Low risk of bias (the study is comparable to a well-performed randomized trial with regard to this domain) • Moderate risk of bias (the study is sound for a non-randomized study with regard to this domain but cannot be considered comparable to a well-performed randomized trial) • Serious risk of bias (the study has some important problems in this domain) • Critical risk of bias (the study is too problematic in this domain to provide	• Definitely low (There is direct evidence of low risk of bias practices) • Probably low (There is indirect evidence of low risk of bias practices OR it is deemed that deviations from low risk of bias practices for these criteria during the study would not appreciably bias results) • Probably high (There is indirect evidence of high risk of bias practices **OR** there is insufficient information (e.g., not reported or	• Low/minimal concerns: Information on the study design and methodologies indicates that they are close to the ideal study characteristics and that the potential for bias is low or minimal, recognizing the general limitations of observational studies. (+++, high quality) • Some concerns: The study design or methodologies are less than ideal, indicating possible bias. (++, medium quality)	• Two stars (can be awarded for Comparability if study controls for two important factors) • One star (meets criteria) • No stars (does not meet criteria) • No description (no stars awarded)

[1] Morgan RL, Thayer K, Santesso N, Holloway AC, Blain R, Eftim S, Goldstone A, Ross P, Guyatt G, Schünemann H: Need for an instrument to evaluate Risk of Bias in Non-randomized Studies of Exposure: Rationale and preliminary instrument. Unpublished.

[2] NTP (National Toxicology Program): Handbook for Conducting a Literature-Based Health Assessment Using Office of Health Assessment and Translation (OHAT) Approach for Systematic Review and Evidence Integration. January 9, 2015 release. Available at http://ntp.niehs.nih.gov/go/38673. 2015.

[3] NTP (National Toxicology Program): Handbook for Preparing Report on Carcinogens Monographs - July 2015. Available at http://ntp.niehs.nih.gov/go/rochandbook. 2015(January 3, 2017).

[4] Wells G, Shea B, O'Connell D, Peterson J, Welch V, Losos M, Tugwell P: The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. Ottawa Hospital Research Institute, 2014. In.: oxford. asp; 2015.

	RoB instrument for NRS of exposures[1]	OHAT[2]*	ORoC[3±]	NOS[4]
	• any useful evidence on the effects of exposure) • No information on which to base a judgement about risk of bias for this domain.	'NR') provided about relevant risk of bias practices) • Definitely high (There is direct evidence of high risk of bias practices)	• Major concerns: The information on the study design or methodologies suggests that the potential for a specific type of bias is high. However, depending on the direction and distortion of the potential bias, the study may have some limited utility. (+, low quality) • Critical concern: Distortion of bias would make the study findings unreliable for cancer hazard identification. ('0' rating) • No information: The information in the study is inadequate to evaluate the level of concern for the domain.	
Study-level responses	• Low risk of bias (the study is comparable to a well-performed randomized trial); • Moderate risk of bias (the study provides sound evidence for a non-randomized study but cannot be considered comparable to a well-performed randomized trial); • Serious risk of bias (the study has some important problems); • Critical risk of bias (the study is too problematic to provide any useful evidence and should not be included in any synthesis); and • No information (No information on which to base a judgement about risk of bias.)	• Tier 1 ('definitely low' or 'probably low' risk of bias for most other applicable criteria) • Tier 2 • Tier 3 (definitely high' or 'probably high' risk of bias for most other applicable criteria)	• High (low/minimal concerns about most potential biases, high or moderate sensitivity rating) • Moderate (low/minimal or some concerns about most potential biases, high or moderate sensitivity rating) • Moderate/low (some or major concerns about several potential biases, sensitivity rating varies) • Low (major concerns about several potential biases, sensitivity rating varies) • Inadequate (critical concerns about any bias, sensitivity rating varies)	• None

	RoB instrument for NRS of exposures[1]	OHAT[2*]	ORoC[3±]	NOS[4]
Outcome-level responses for body of evidence	• Not serious • Serious • Very serious • Critically serious	• High • Moderate • Low • Very Low	• None	• None

*Considered to be similar to Navigation Guide Instrument; ±Considered to be similar to EPA-IRIS instrument. NOS: Newcastle-Ottawa Scale; NRS: non-randomized studies; OHAT: Office of Health Assessment and Technology; ORoC: Office on the Report on Carcinogens; RoB: risk of bias.

Appendix B. Topic-specific expert observations per topic area and study.

Topic area	Study	Observations (n)
PFOA and fetal growth	Apelberg et al., 2007	6
	Hamm et al., 2010	6
	Kim et al., 2011	6
	Maisonet et al., 2012	6
	Nolan et al., 2009	6
	Whitworth et al., 2012	6
BPA and weight	Bhandari et al., 2013	8
	Carwile & Michels, 2011	8
	Harley et al., 2013	8
	Shankar et al., 2012	8
	Wang et al., 2012	8
	Zhao et al., 2012	8
PBDE and thyroid simulation hormone	Cao et al., 2011	4
	Cesaroni et al., 2013	4
	Hystad et al., 2013	4
	Katanoda et al., 2011	4
	Krewski et al., 2009	4
	Lepeule et al., 2012	4
PM2.5 and lung cancer	Bloom et al., 2008	5
	Han et al., 2010	5
	Kim et al., 2012	3
	Kim et al., 2013	5
	Lin et al., 2011	5
	Stapleton et al., 2011	5
Phthalates and pre-term birth	Adibi et al., 2009	5
	Meeker et al., 2009	4
	Suzuki et al., 2010	5
	Whyatt et al., 2009	5
	Wolff et al., 2008	5

Appendix C. Construct validity: Pearson correlation coefficients across similar instrument domains.

Domains

Confounding					
		RoB instrument for NRS of exposures	OHAT	ORoC	NOS
1	RoB instrument for NRS of exposures	1	0.69	0.62	0.79
2	OHAT		1	0.88†	0.62
3	ORoC			1	0.58
4	NOS				1

Selection					
		RoB instrument for NRS of exposures	OHAT	ORoC	NOS
1	RoB instrument for NRS of exposures	1	0.63†	0.41	0.37
2	OHAT		1	0.81	0.46
3	ORoC			1	0.54
4	NOS				1

Exposure					
		RoB instrument for NRS of exposures	OHAT‡	ORoC	NOS
1	RoB instrument for NRS of exposures	1	0.92	0.92	0.43
2	OHAT‡		1	0.95	0.38
3	ORoC			1	0.31*
4	NOS				1

Departures from intended exposure					
		RoB instrument for NRS of exposures	OHAT	ORoC**	NOS
1	RoB instrument for NRS of exposures	1	x	0.49	x
2	OHAT		1	x	x
3	ORoC			1	x
4	NOS				1

Missing data					
		RoB instrument for NRS of exposures	OHAT	ORoC	NOS
1	RoB instrument for NRS of exposures	1	0.61†	0.29	x
2	OHAT		1	0.26	x
3	ORoC			1	x
4	NOS				1

Outcome					
		RoB instrument for NRS of exposures	OHAT§	ORoC	NOS
1	RoB instrument for NRS of exposures	1	0.52	0.58	0.22*
2	OHAT§		1	0.67	0.11*
3	ORoC			1	0.20*
4	NOS				1

Reporting					
		RoB instrument for NRS of exposures	OHAT	ORoC	NOS
1	RoB instrument for NRS of exposures	1	0.23*	0.36	x
2	OHAT		1	0.47	x
3	ORoC			1	x
4	NOS				1

Other^					
		RoB instrument for NRS of exposures	OHAT	ORoC	NOS
1	RoB instrument for NRS of exposures	1	x	x	x
2	OHAT		1	0.58	x
3	ORoC			1	x
4	NOS				1

* = not significant at $p < 0.05$; x = no comparable concept measured; ^ = measures the other sources of bias domain in OHAT and the study sensitivity domain in ORoC; † = scatterplots suggested potentially spurious relationships; ‡ = measures the detection of exposures bias domain in OHAT; ** = measures the analysis domain in ORoC; § =

measures the detection of outcomes bias domain in OHAT. NOS: Newcastle-Ottawa Scale; NRS: non-randomized studies; OHAT: Office of Health Assessment and Technology; ORoC: Office on the Report on Carcinogens; RoB: risk of bias.

Appendix D. Domains demonstrating moderate or high Pearson correlation coefficients with other domains within the same instrument.

	Domain	Domains in same instrument suggesting moderate or strong correlations*	Pearson correlation coefficient (*r*)
RoB instrument for NRS of exposures	Bias due to selection of participants	Bias due to confounding	0.51
		Bias due to intended exposures	0.64
		Bias in measurement of outcomes	0.63
OHAT	Selection bias	Confounding bias	0.51
		Detection of outcomes bias	0.54
		Selective reporting bias	0.52
		Other sources of bias	0.58
	Confounding bias	Selective reporting bias	0.53
		Other sources of bias	0.51
	Detection of outcomes bias	Selective reporting bias	0.51
	Selective reporting bias	Other sources of bias	0.56
ORoC	Selection bias	Confounding	0.50
		Selective reporting	0.53
		Analysis	0.59
	Outcome	Analysis	0.77
		Study sensitivity	0.61
	Analysis	Confounding	0.57
		Study sensitivity	0.66
NOS	Selection bias	Comparability	0.57

* = significance measured at $p < 0.05$. NOS: Newcastle-Ottawa Scale; NRS: non-randomized studies; OHAT: Office of Health Assessment and Technology; ORoC: Office on the Report on Carcinogens; RoB: risk of bias.

Appendix E. Domains demonstrating moderate or high Pearson correlation coefficients with other domains in different instruments.

	Domain	Domain in different instrument suggesting moderate or strong correlations*	Pearson correlation coefficient (*r*)
RoB instrument for NRS of exposures	Bias due to confounding	OHAT selection bias	0.69
		ORoC selection bias	0.64
		ORoC outcome measures	0.53
		ORoC analysis	0.71
	Bias in selection of participants	OHAT selective reporting bias	0.53
		ORoC outcome measures	0.51
		ORoC analysis	0.54
	Bias due to missing data	OHAT outcome measures	0.50
		OHAT other sources of bias	0.52
		ORoC reporting	0.79
	Bias in measurement of outcomes	OHAT selective reporting bias	0.58
		ORoC analysis	0.53
	Bias in reported results	ORoC confounding bias	0.51
OHAT	Selection bias	RoB instrument for NRS of exposures bias due to confounding	0.69
		ORoC outcome measures	0.55
		ORoC confounding	0.54
		ORoC analysis	0.64
		NOS comparability	0.53
	Confounding bias	ORoC selection bias	0.56
		ORoC analysis	0.62
	Performance bias	ORoC selection bias	0.54
		ORoC reporting	0.59
	Detection of outcomes bias	RoB instrument for NRS of exposures bias due to missing data	0.50

		ORoC analysis	0.61
	Selective reporting bias	RoB instrument for NRS of exposures bias in selection of participants	0.53
		RoB instrument for NRS of exposures bias due to the measurement of outcomes	0.58
		ORoC selection bias	0.51
		ORoC outcome measures	0.57
		ORoC confounding	0.56
		ORoC analysis	0.69
	Other sources of bias	RoB instrument for NRS of exposures bias due to missing data	0.52
		ORoC selection bias	0.54
		ORoC outcome measures	0.73
		ORoC reporting	0.57
		ORoC analysis	0.67
ORoC	Selection bias	RoB instrument for NRS of exposures bias due to confounding	0.64
		RoB instrument for NRS of exposures bias due to missing data	0.57
		OHAT confounding bias	0.56
		OHAT performance bias	0.54
		OHAT selective reporting bias	0.51
		OHAT other sources of bias	0.54
		NOS comparability	0.61
	Outcome	RoB instrument for NRS of exposures bias due to confounding	0.53
		RoB instrument for NRS of exposures bias due to selection of participants	0.51
		OHAT selection bias	0.55

		OHAT selective reporting bias	0.57
		OHAT other sources of bias	0.73
	Confounding	RoB instrument for NRS of exposures bias in reported results	0.51
		OHAT selection bias	0.54
		OHAT reporting bias	0.56
	Reporting	RoB instrument for NRS of exposures bias due to missing data	0.79
		OHAT performance bias	0.59
		OHAT other sources of bias	0.57
	Analysis	RoB instrument for NRS of exposures bias due to confounding	0.71
		RoB instrument for NRS of exposures bias due to selection of participants	0.54
		RoB instrument for NRS of exposures bias in the measurement of outcomes	0.53
		OHAT selection bias	0.64
		OHAT confounding bias	0.62
		OHAT detection of outcomes bias	0.61
		OHAT selective reporting bias	0.69
		OHAT other sources of bias	0.67
		NOS selection	0.53
		NOS comparability	0.56
	Study sensitivity	OHAT selective reporting bias	0.50
		NOS selection	0.73
		NOS comparability	0.54
NOS	Selection	ORoC analysis	0.53
		ORoC study sensitivity	0.73
	Comparability	OHAT selection bias	0.53

		ORoC selection	0.61
		ORoC analysis	0.56
		ORoC study sensitivity	0.54

* = significance measured at $p < 0.05$. NOS: Newcastle-Ottawa Scale; NRS: non-randomized studies; OHAT: Office of Health Assessment and Technology; ORoC: Office on the Report on Carcinogens; RoB: risk of bias.

Appendix F. Results from mean time-burden comparison analysis.

Between instrument analysis of variance (ANOVA) for the variable time, as measured in minutes: $p < 0.001$

Multiple comparisons across instruments and ratings by topic-specific experts using Tukey's Honest Significant Difference (HSD) test: mean difference (p-value)

Instrument	RoB instrument for NRS of exposures	OHAT	ORoC	NOS	Topic-specific experts*
RoB instrument for NRS of exposures	x	39.79 (p<0.001)	47.41 (p<0.001)	67.12 (p<0.001)	30.63 (p<0.001)
OHAT		x	7.62 (p=0.008)	27.33 (p<0.001)	9.16 (p=0.002)
ORoC			x	19.71 (p<0.001)	16.78 (p<0.001)
NOS				x	36.50 (p<0.001)
Topic-specific experts*					x

The mean difference is significant at the 0.05 level. NOS: Newcastle-Ottawa Scale; NRS: non-randomized studies; OHAT: Office of Health Assessment and Technology; ORoC: Office on the Report on Carcinogens; RoB: risk of bias. *Based on 160 observations across 29 studies.

CHAPTER 6. CONCLUSIONS

1. Summary of findings

This work presents four main pieces of research. The main findings can be summarized as the following:

1. A structured framework is needed to facilitate decision-making in the environmental health field and, given the successful application of the GRADE framework in many other disciplines, I suggest this approach be explored.
2. To harmonize the methodological and environmental health decision-making process, evaluation is needed for the following priority areas: research question formulation, evaluating the certainty of evidence, and making recommendations.
3. When identifying questions about environmental health hazards, five strategies can be used to facilitate identification of the exposure and comparator: 1) use the cut-offs achieved from implementation of existing interventions; 2) use existing exposure cut-offs (e.g., thresholds, levels, ranges, or durations) associated with the known health outcomes of interest; 3) when only the exposure for a population is knows, use mean thresholds from external or general populations (from other research); 4) use cut-offs defined based on the distribution in studies identified from a search or scoping review; and 5) use the distribution of the relationship between the exposure and outcome.
4. Evaluating the certainty of evidence requires an assessment of the risk of bias (RoB) of individual studies. The RoB instrument for non-randomized studies (NRS) of exposures can be used to evaluate individual studies and provide

judgments across a body of evidence for environmental health exposures by using the concept of the target experiment as a point of reference.

5. The integration of the RoB instrument for NRS of exposures into the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) framework allows for NRS to start at 'High' initial certainty of evidence (CoE), instead of the default initial certainty of 'Low'.
6. Reliability and validity testing of the RoB instrument for NRS of exposures revealed robust measurement properties meaning that this instrument can be used by wider audiences.

2. Reflections of an effort to develop a standardized instrument to evaluate risk of bias in studies of exposure and implications for decision making

2.1. Inception

There are factors that can reduce the uncertainty in a body of evidence and improve the accuracy of decision making for human health outcomes. In this work, I recognized a need for further methodological development in the evaluation of RoB and integration into decision-making for studies of exposures [1]. My objective was to develop a RoB instrument to assess bias in NRS of exposures and to prompt researchers to improve the study design of future studies. The implications of this instrument are that it can be used not only for policy and decision-making, but also to inform future research. I approached this process in four stages:

1) Exploration and recognition of current practices for decision-making in environmental health and identification of priority areas for further methods research;
2) Pilot testing of a Risk of Bias Instrument for Non-randomized Studies of Interventions (ROBINS-I) and modifications to tailor for NRS of exposures;
3) Application of the RoB instrument for NRS of exposures;
4) Integration of the RoB instrument for NRS of exposures into the GRADE approach; and
5)
6) Reliability and validity testing of the RoB instrument for NRS of exposures and comparison with other RoB instrument used for studies of exposures and topic-specific expert evaluation of studies.

In 2016, ROBINS-I was released to evaluate RoB within studies of health interventions [2]. ROBINS-I combined several distinct concepts for the evaluation of studies of interventions. First, ROBINS-I introduced the evaluation of NRS as a standardized comparison to randomized-controlled trials (RCTs) by using an absolute scale to measure bias [2, 3]. First, implications of this design include the possibility to evaluate NRS like RCTs on a bias domain level against the possible least biased (hypothetical) randomized trial. Although final RoB ratings across studies should not differ, it avoids a two-step approach of rating RoB in GRADE twice and it provides a more nuanced assessment of confounding and selection bias. In particular, the instrument highlights domains of bias distinct to NRS when randomization and allocation concealment are not

part of the experiment, essentially what had lead developers to start NRS at 'Low' within GRADE initially. Therefore, NRS and RCTs start at the same initial level of a 'High' certainty rating. Second, ROBINS-I instructed users to incorporate the hypothesized direction of bias into their judgements at the domain and final study-level RoB. The implication is that bias is not a clear-cut issue and deeper understanding of how much the bias is expected to modify the effect estimate in the analysis from the true estimate may lead a rater to have greater concerns about the introduction of bias (i.e., the bias overestimates the effect of an intervention) or less concerns about the introduction of bias (i.e., the introduced bias is more conservative).

I decided to examine the potential for ROBINS-I within the environmental health field based on feedback from systematic-review authors and guideline developers in the field. Many preferred the concept of transparently recognizing that NRS suffer from substantial bias due to the lack of a balance of prognostic factors within a RoB instrument instead of the automatic start at 'Low' within GRADE. However, pilot testing of ROBINS-I revealed some conceptual and semantic modifications leading to the adaptation of the instrument for studies of exposures (i.e., ROBINS of Exposures) [4]. These modifications included replacement of the term 'intervention' with 'exposure' throughout the instrument; renaming of 'target trial' to 'target experiment' and broadening the definition to include animal experiments; the addition of fields in the preliminary stages of the instrument to collect information on the accuracy of measurement of exposures and outcomes to guide the rater to distinguish between issues of indirectness and risk of bias (RoB); and inclusion of additional signaling

questions to assess bias in exposure measurement. While the ROBINS-I instrument identified the domains of confounding and selection of participants to related to the loss of randomization and allocation concealment, I recognized that bias due to misclassification of the exposure may also result from the lack of a prognostic balance, namely the ability to correctly distinguish between the exposure and comparison of interest.

I recognized a paucity of research exploring and establishing the reliability and validity of current instruments used to evaluate RoB within studies of exposures. To understand and aid in the development of this instrument, I evaluated the interrater reliability of the RoB instrument for NRS of exposures. In addition, I compared the interrater and inter-instrument reliability across three other instruments commonly used in the field. I determined the RoB instrument for NRS of exposures to have substantial reliability and be ready for wider use.

As no reference standard has been established for RoB evaluation of exposure studies, I compared the correlation coefficients of the RoB instrument for NRS of exposures with three other instruments used in the field, and unstructured evaluations from exposure topic-specific experts. In addition, I determined the construct validity of the RoB instrument for NRS of exposures to be in strong agreement with other instruments used to evaluate exposure studies and evaluations by topic-specific experts.

A recently submitted guidance document summarizes the integration of ROBINS-I into GRADE and the implication of using an absolute scale for the RoB assessment of RCTs and NRS [3]. Key points include that when using a standardized comparison, all studies

would start at 'High' initial certainty within GRADE; based on concerns of confounding and selection bias, NRS would be expected to end up as at least very serious RoB within GRADE (equivalent to starting at 'Low' initial certainty); and that the factors considered to rate up NRS (magnitude of effect, dose response, or opposable residual confounding) could be considered during the RoB assessment and inform the RoB judgment within GRADE. While similar for the integration of the RoB instrument for NRS of exposures into GRADE, I recognize a few distinguishing concepts: 1) strategies for identifying an exposure and comparison within the research question; 2) the prominence that misclassification of exposure has on the overall RoB judgment; 3) an algorithm outlining the process within GRADE for using this RoB instrument for NRS of exposures; and 4) the distinction between sources of bias and indirectness. Recognizing that identifying a specific exposure threshold (e.g., levels, durations, ranges, means, medians, or ranges of exposure) is difficult when little information exists categorizing an exposure as a harm or the definition of an exposure, I propose that there are five strategies for identifying the threshold(s) of interest: 1) use the cut-offs achieved from implementation of existing interventions; 2) use existing exposure cut-offs (e.g., thresholds, levels, ranges, or durations) associated with the known health outcomes of interest; 3) when only the exposure for a population is knows, use mean thresholds from external or general populations (from other research); 4) use cut-offs defined based on the distribution in studies identified from a search or scoping review; and 5) use the distribution of the relationship between the exposure and outcome [5]. When applying ROBINS-I, studies would typically be judged as 'serious' RoB following the evaluation of confounding and

selection bias. Within the RoB instrument for NRS of exposures, misclassification of the exposure would also be expected to typically lead studies to a 'serious' judgment because of the difficulty in appropriately classifying unintentional exposures. Finally, to facilitate implementation of this RoB instrument for NRS of exposures, I provide an algorithm to guide reviewers through the process. This algorithm highlights actions throughout the process and how the RoB instrument for NRS of exposures judgments integrate into GRADE (Figure 1). As demonstrated within the algorithm, the RoB instrument for NRS of exposures also recognizes the common conflation between issues of bias and issues of indirectness (i.e., generalizability or applicability), external to RoB. Both constructs inform the GRADE evidence assessment; however, indirectness should not be included as rationale when making a judgment of RoB.

My project highlights the value added by incorporating the RoB instrument for NRS of exposures for evaluation of individual studies within systematic reviews and GRADE to inform decision-making about environmental health. The algorithm outlining the steps within this RoB instrument for NRS of exposures and examples detailing the application of the instrument and the integration into GRADE should facilitate use of the instrument. Users should feel more confident using the instrument based on the results of the reliability and validity study. In addition, this work highlights concerns and solutions for systematic-review authors and guideline developer when answering questions about environmental exposures. This work also highlights areas of importance for researchers developing primary studies of exposures. Implications are that

researchers can improve their research by addressing those areas and common domains, thus reducing uncertainty about the effect estimate within their study.

2.2. Challenges during the process

The motivation behind this project came as an attempt to harmonize among different environmental-health organizations and to harmonize environmental-health methods with current systematic review and guideline methodologies, specifically the GRADE approach. While a straightforward objective on paper, this has been one of the most challenging issues throughout the project, revealing many systemic issues. Not only are there semantic differences between disciplines (e.g. the terms 'study sensitivity' and 'sensitivity analysis'), there are also a wide breadth of exposure topics that one instrument hopes to address. This leads to frustration in both disciplines. For the methodologists, can the fidelity of the methods be maintained? And among the environmental health field, can these methods be pragmatically implemented and understood? How do we keep the fidelity and rigor of the process but make the process useful and more desirable than other options? While I recognize the presence of instruments deemed acceptable and used by environmental-health organizations to evaluate RoB, if there wasn't some discontent than this project would not have gained traction.

The process of adapting ROBINS-I for studies of exposures identified a few areas of discontent. Feedback from developers of the ROBINS-I instrument on the modified instrument, this RoB instrument for NRS of exposures, came from two extremes: 1) ROBINS-I should be applicable to studies of exposure without changes; and 2) if these

modifications to ROBINS-I are needed, then why not just modify ROBINS-I instead of making a new instrument. Justifying why ROBINS-I required modifications for studies of exposure becomes a struggle, if one denies that even semantic adaptations are unnecessary for acceptance and adoption. Our first pilot test results demonstrated that the use of the word 'intervention' alienated raters to the point of misunderstanding and indifference to the instrument. Replacement with the term 'exposure', even though the meaning in this situation was essentially the same, improved understanding and application of the instrument. The semantic modification of replacement of that term, while regarded as superficial, led to greater acceptance of the instrument. Just as ROBINS-I requires more granularity in the assessment of RoB items than the Cochrane RoB tool for RCTs, the RoB instrument for NRS of exposures requires more granularity to evaluate studies of exposures.

One desirable attribute of ROBINS-I identified from environmental-health scientists was the ability for studies of all designs to start at 'High' initial certainty within the GRADE framework. By starting studies of all designs at 'High' initial certainty does not mean that they are all devoid of bias. In fact, the potential for study limitations within NRS comes under greater scrutiny, as NRS are held to the standards of well-conducted RCTs. The typical RoB judgment becomes that studies are recognized as potentially introducing 'Serious' RoB, unless exploration of residual and unmeasured confounding identifies that the reported effect estimate most likely does not deviate much from the true estimate. Depending on the exposure of interest, information evaluating the potential for confounding is limited.

Further exploration of the RoB instrument for NRS of exposures presented a greater challenge, when compared with other commonly used instruments to evaluate NRS of exposures. As presented in Chapter 5, interrater and inter-instrument reliability is comparable across the RoB instrument for NRS of exposures and three other commonly used instruments: Newcastle-Ottawa Scale (NOS), and tools used by the National Toxicology Programs' Office of Health Assessment and Translation (OHAT), and Office of the Report of Carcinogens (ORoC). However, the RoB instrument for NRS of exposures takes significantly longer to complete than the alternative instruments. The results from the construct validity analyses conducted at the study and domain levels revealed the potential of the RoB instrument for NRS of exposure to facilitate the evaluation process, even among users without specific expertise in the exposure of interest. Therefore, while there is indeed a trade-off considering the time investment, the validity analyses suggest that the concepts measured in the domains of the RoB instrument for NRS of exposure are discrete and more explicit than those in the other instruments.

2.3. Next steps

This work presents a few of the many methodological advancements in the field of environmental-health decision-making. I recognize that the RoB instrument for NRS of exposures represents a preliminary instrument for study evaluation. Currently, efforts are underway to broaden the methodological and environmental health input on the RoB instrument for NRS of exposure to develop a ROBINS of exposures (ROBINS-E) instrument, to complement ROBINS-I. This initiative has increased the type of exposures

(i.e. to include occupational exposures) and number of applications of the questions to see if adaptations to the RoB instrument for NRS of exposures are needed.

NRS represent many but not all study types available to inform decision-making about environmental exposures. I focus on how NRS can be evaluated using standardized and transparent methods, but that is only one piece. Researchers have identified current practices of organizations using both NRS and RCT evidence to inform recommendations, presenting suggestions for how to integrate the two evidence streams. Research is on-going for ways to evaluate RoB and assess the certainty of studies of animal, in vitro, and mechanistic evidence. Additionally, initiatives are addressing how to present multiple evidence streams in a standardized way.

Collectively, this research should move us forward in the field; however, there is still much to do. In any political climate or financial situation, recommendations and policies should be held accountable to the underlying evidence. In this work, we attempt to explore methods for understanding and bringing transparency to that evidence; however, further evaluation of this work is needed. The acceptance and adoption of new methods or modifications to current practices require a comprehensive approach to behavior change at the individual and societal level. Maintaining communication between methodologists and environmental health scientists may increase opportunities for evaluation, feedback, and further development.

2.4. Final thoughts

"There is high demand in environmental health for adoption of a structured process that evaluates and integrates evidence while making decisions and recommendations transparent" [1]. We explored the possibility of the adoption of methods for evidence assessment and decision-making in the environmental-health field; however, determined that modifications were needed to address our objectives. When evaluating the certainty of the evidence, instruments should adapt to intricacies within the environmental exposures literature. As with any novel instrument, exploration is needed to understand the robustness of the instrument and that it performs in the hypothesized way, which includes application to a variety of exposures and study designs. The RoB instrument for NRS of exposures represents the product of a multi-stage development process reflective of these considerations. This instrument facilitates a structured evaluation of exposure studies to inform decision-making. Wider adoption of these methods will reveal areas for further development.

3. References

1. Morgan RL, Thayer KA, Bero L, Bruce N, Falck-Ytter Y, Ghersi D, Guyatt G, Hooijmans C, Langendam M, Mandrioli D *et al*: **GRADE: Assessing the quality of evidence in environmental and occupational health**. *Environ Int* 2016, **92-93**:611-616.
2. Sterne JA, Hernan MA, Reeves BC, Savovic J, Berkman ND, Viswanathan M, Henry D, Altman DG, Ansari MT, Boutron I *et al*: **ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions**. *BMJ* 2016, **355**:i4919.
3. Schünemann H, Cuello C, Akl EA, Mustafa R, Meerpohl J, Thayer K, Morgan R, Gartlehner G, Kunz R, Katikireddi S *et al*: **GRADE Guidelines: 18. How tools to assess risk of bias in non-randomized studies should be used to rate the certainty of a body of evidence**. Unpublished.
4. Morgan R, Thayer K, Holloway A, Santesso N, Blain R, Eftim S, Goldstone A, Ross P, Guyatt G, Schünemann H: **Application of a Risk of Bias in Non-randomized Studies of Exposure (ROBINS for exposures) instrument and integration into certainty in the evidence assessments using GRADE**. Unpublished.
5. Morgan RL, Thayer K, Whaley P, Schünemann H: **Identifying the PECO: A framework for formulating good questions to explore the association of environmental and other exposures with health outcomes**. Unpublished.

Figures

Figure 1.

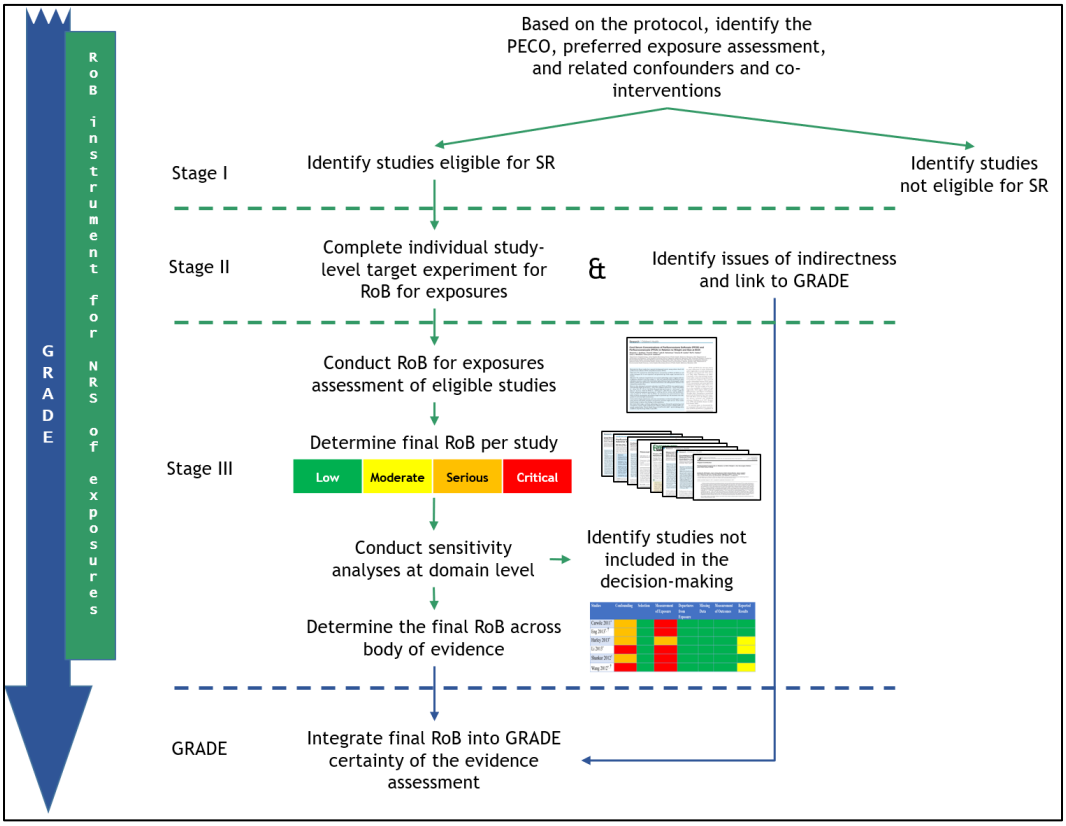


Figure 1. Approach for conducting an assessment using the RoB instrument for NRS of exposures and the integration into GRADE when conducting systematic reviews of exposure. From "Risk of Bias instrument for Non-randomized Studies of exposures: a users' guide".